Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Πολυτροπική Αναγνώριση Χειρονομιών

Διπλωματική Εργασία

του

Γεωργίου Η. Παυλάκου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2014

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Πολυτροπική Αναγνώριση Χειρονομιών

Διπλωματική Εργασία

του

Γεωργίου Η. Παυλάκου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 16η Ιουλίου 2014.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Πέτρος Μαραγκός	Γεώργιος Παπαβασιλόπουλος	Κωνσταντίνος Τζαφέστας
Καθηγητής	Καθηγητής	Επίκουρος Καθηγητής
Ε.Μ.Π.	Ε.Μ.Π.	Ε.Μ.Π.

Αθήνα, Ιούλιος 2014

(Υπογραφή)

.....
Γεώργιος Η. Παυλάκος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Κατ΄ αρχάς, θα ήθελα να ευχαριστήσω θερμότατα τον καθηγητή κ. Πέτρο Μαραγκό, υπό την επίβλεψη του οποίου εκπονήθηκε η παρούσα διπλωματική. Ήταν μεγάλη χαρά και τιμή για μένα που συνεργάστηκα μαζί του, και του χρωστάω πολλά για όλες τις ευκαιρίες που μου έδωσε σε αυτό το διάστημα. Επίσης, αισθάνομαι ευγνώμων απέναντι στους Σταύρο Θεοδωράκη και Βασίλη Πιτσικάλη, για τη στενή καθοδήγηση και τη βοήθειά τους σε αυτή μου την προσπάθεια. Αντίστοιχα, ευχαριστώ και όλα τα μέλη του εργαστηρίου Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων, για την υποστήριξή τους οποτεδήποτε χρειάστηκε, και ειδικά τους Νάσο Κατσαμάνη, Πέτρο Κούτρα και Κέβη Μανίνη. Επίσης, στα πλαίσια της ερευνητικής συνεργασίας μου με το εργαστήριο Ρομποτικής και Αυτοματισμού, ευχαριστώ ιδιαίτερα τον επίκουρο καθηγητή κ. Κώστα Τζαφέστα καθώς και τις Ξανθή Παπαγεωργίου και Γεωργία Χαλβατζάκη. Ακόμα, ιδιαίτερη αναφορά θέλω να κάνω στους συμφοιτητές μου Βαγγέλη, Γιώργο, Γρηγόρη και Μάγια, για τη συνεργασία που είχαμε στο πρώτο διάστημα της διπλωματικής μου, καθώς και όλους τους συμφοιτητές και φίλους που είχα τη χαρά να συνεργαστώ και να αλληλεπιδράσω κατά τη διάρκεια των προπτυχιακών μου σπουδών. Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, για τη στήριξή τους, για την κατανόησή τους, για όσα μου έχουν προσφέρει όλα αυτά τα χρόνια.

Περίληψη

Η συγκεκριμένη διπλωματική έχει σαν αντικείμενο την αντιμετώπιση του προβλήματος της αναγνώρισης χειρονομιών, και των τεχνικών πολυτροπικής σύμμειξης που μπορούν να εφαρμοστούν. Μελετάται η μοντελοποίηση και η αναγνώριση των χειρονομιών με χρήση ισχυρών εργαλείων όπως τα Κρυφά Μαρκοβιανά Μοντέλα, αλλά και άλλων ταξινομητών μηχανικής μάθησης, όπως τα Support Vector Machines και k-Nearest Neighbor. Για την εξαγωγή χαρακτηριστικών χρησιμοποιούμε το κανάλι πληροφορίας της χειρομορφής, από όπου εξάγουμε δημοφιλείς οπτικούς περιγραφητές, όπως τα Histograms of Oriented Gradients (HOG), αλλά και το κανάλι πληροφορίας της θέσης-κίνησης, όπου τα χαρακτηριστικά προκύπτουν από τη θέση (σχετική θέση, απόσταση) και την κίνηση (ταχύτητα, διεύθυνση), του χεριού και του αγκώνα. Τέλος, παρουσιάζουμε δύο επιτυχημένα σχήματα σύμμειξης αυτών των δύο καναλιών οπτικής πληροφορίας με την τροπικότητα του ήχου. Μάλιστα, τα αποτελέσματά μας σε πολυτροπική βάση αναγνώρισης χειρονομιών, ξεπερνούν τις επιδόσεις που επιτεύχθηκαν σε πρόσφατο διαγωνισμό πολυτροπικής αναγνώρισης χειρονομιών.

Λέξεις Κλειδιά

όραση υπολογιστών, πολυτροπική αναγνώριση χειρονομιών, επικοινωνία ανθρώπου-υπολογιστή, αισθητήρας Kinect, κρυφά Μαρκοβιανά μοντέλα, ιστογράμματα προσανατολισμένων gradients, σχήματα πολυτροπικής σύμμειξης

Abstract

This thesis focuses on the gesture recognition problem and on multimodal fusion techniques for it. We study gesture modeling and recognition using powerful tools, such as Hidden Markov Models, as well as other machine learning classifiers, like Support Vector Machines and K-Nearest Neighbor. For feature extraction we focus on Handshape information, employing various visual descriptors, like Histograms of Oriented Gradients (HOG), and Movement-Position information, where features are extracted based on the position (relative position, distance) and the movement (velocity, direction) of hands and elbows. Finally, we present two successful fusion schemes, employing both visual cues and audio modality. Our proposed methodology achieves high gesture recognition accuracy in a multimodal gesture dataset, outperforming all recently published approaches on the same challenging gesture recognition task.

Keywords

computer vision, multimodal gesture recognition, human-computer interaction, Kinect sensor, hidden Markov models, histograms of oriented gradients, multimodal fusion schemes

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Το πρόβλημα της Αυτόματης Αναγνώρισης Χειρονομιών και η σημασία του

Το πρόβλημα της Αναγνώρισης Χειρονομιών (Gesture Recognition) αποτελεί ερευνητικό πεδίο της Όρασης Υπολογιστών και της Αναγνώρισης Προτύπων, που σκοπό έχει την ανάπτυξη αλγορίθμων και μεθόδων για την αυτόματη επεξεργασία της οπτικής πληροφορίας και της αναγνώρισης των χειρονομιών σε αυτή. Στη βιβλιογραφία, ο όρος gesture αφορά γενικότερες κινήσεις που σκοπό έχουν να μεταδώσουν κάποιο μήνυμα, είτε πρόκειται για κινήσεις των χεριών, όπως συμβαίνει συνήθως, είτε για κινήσεις του προσώπου, του κεφαλιού, ή ακόμα και του σώματος [53]. Ωστόσο, στα πλαίσια της παρούσας διπλωματικής με τον όρο χειρονομία θα αναφερόμαστε κυρίως στις κινήσεις των χεριών, ακόμα και αν αυτές συνοδεύονται από αντίστοιχες εκφράσεις του προσώπου, κινήσεις του κεφαλιού, ή αλλαγές της στάσης.

Ο λόγος που ένα πρόβλημα σαν αυτό της αναγνώρισης χειρονομιών παρουσιάζει τέτοιο ενδιαφέρον στην ερευνητική κοινότητα, είναι η θέση που κατέχουν οι χειρονομίες στην επικοινωνία μας. Πολύ συχνά συνοδεύουν το λόγο μας, μπορεί να χρησιμοποιηθούν για να εκφράσουν τα συναισθήματά μας, ή αρκούν για να επικοινωνήσουμε χωρίς τη χρήση ομιλίας. Η φυσικότητα της έκφρασης μέσω των χειρονομιών γίνεται αντιληπτή αν αναλογιστούμε ότι ακόμα και οι εκ γενετής τυφλοί άνθρωποι χειρονομούν κατά τη διάρκεια της ομιλίας τους, σύμφωνα με το άρθρο των Iverson και Goldin-Meadow [36]. Συνεπώς, οι χειρονομίες δεν πρόκειται απλά για ένα πολιτιστικό φαινόμενο στο οποίο προσαρμοζόμαστε με βάση την παρατήρηση, αλλά αυτές προκύπτουν πολύ φυσιολογικά για τους ανθρώπους.

Αυτή η φυσικότητα της επικοινωνίας μέσω χειρονομιών είναι που έχει δημιουργήσει την έμπνευση για σειρά εφαρμογών, που θα μπορούσαν να επωφεληθούν από την ερευνητική πρόοδο στην αυτόματη αναγνώριση χειρονομιών.

Μεταξύ αυτών, ενδεικτικά αναφέρονται:

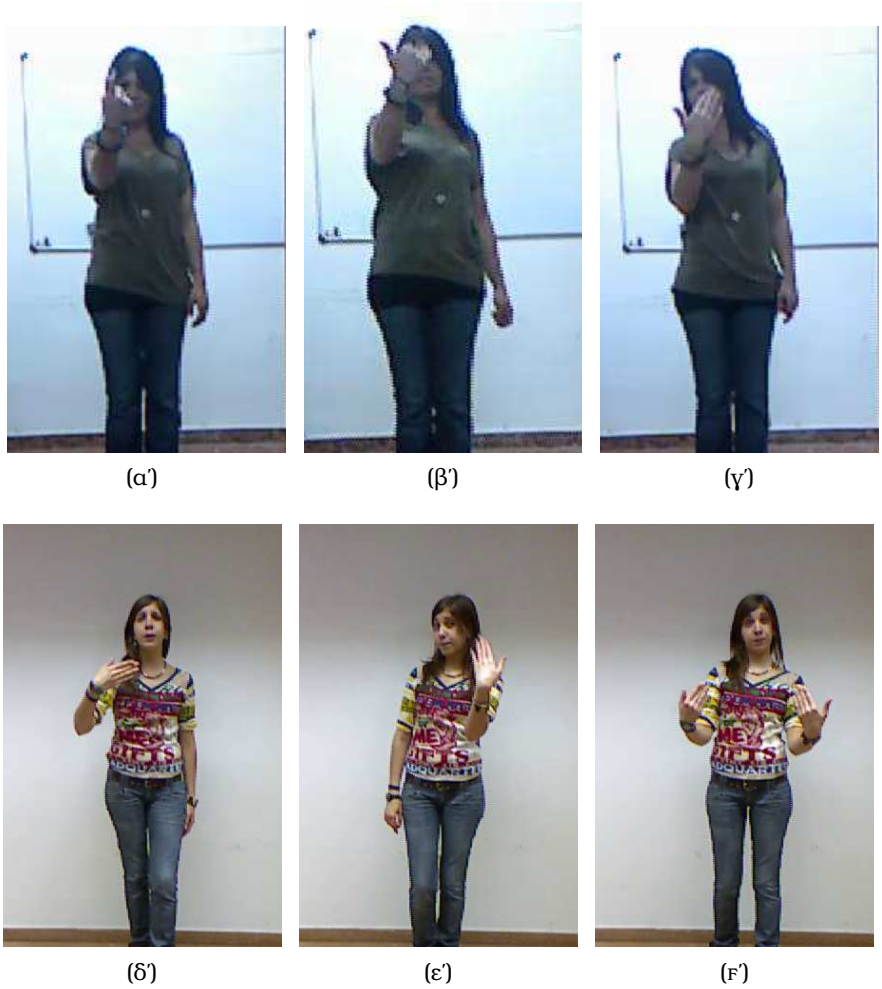
- η ανάπτυξη συστημάτων επικοινωνίας για άτομα με προβλήματα ακοής.
- η διευκόλυνση πρώιμης επικοινωνίας με τον υπολογιστή για παιδιά μικρής ηλικίας.
- η αναγνώριση νοηματικής γλώσσας.
- η πλοήγηση σε περιβάλλοντα εικονικής πραγματικότητας.
- οι εφαρμογές επικοινωνίας σε τηλεδιασκέψεις.

Παρατηρώντας το εύρος των πιθανών εφαρμογών, που φιλοδοξούν από το να βελτιώσουν καθημερινές δραστηριότητες (όπως η διαχείριση μιας τηλεδιάσκεψης) μέχρι και να καλύψουν ουσιαστικές ελλείψεις για κάποιες κοινωνικές ομάδες (όπως η επικοινωνία μέσω νοηματικής γλώσσας), αντιλαμβανόμαστε το διεπιστημονικό ενδιαφέρον του συγκεκριμένου προβλήματος.

Το όραμα συνεπώς σε αυτή την κατεύθυνση, είναι η κατασκευή ενός εύρωστου συστήματος χειρονομιών, το οποίο να ανταποκρίνεται στις οδηγίες διαφορετικών χρηστών. Και μπορεί στο πεδίο της αυτόματης αναγνώρισης ομιλίας (ASR), της έταιρης δηλαδή σημαντικής τροπικότητας που χρησιμοποιούμε για την επικοινωνία μας, η έρευνα να έχει κάνει σημαντική πρόοδο στην ανάπτυξη αντίστοιχων συστημάτων [33], εντούτοις, οι προκλήσεις για ένα οπτικό σύστημα αναγνώρισης χειρονομιών παραμένουν πολλές. Η ποικιλία των χρησιμοποιούμενων χειρομορφών, η σημαντική επίδραση του σημείου λήψης στην οπτική πληροφορία, οι συχνές επικαλύψεις, καθώς και η διακύμανση στον τρόπο εκτέλεσης των χειρονομιών αυξάνουν σημαντικά την πολυπλοκότητα του προβλήματος, και απαιτούν ειδικούς χειρισμούς για την αντιμετώπισή του.

Ειδικά αυτή η διαφοροποίηση στον τρόπο εκτέλεσης, αποτελεί μία από τις σημαντικότερες προκλήσεις του πεδίου, μιας και οι χειρονομίες δεν είναι συνδεδεμένες, με τα αντίστοιχα νοήματα που έχουν σκοπό να εκφράσουν, με αμφιμονοσήμαντο τρόπο. Αντίθετα, υπάρχουν πολλές διακυμάνσεις στον τρόπο εκτέλεσης, μεταξύ διαφορετικών χρηστών, ή ακόμα και μεταξύ επαναλήψεων του ίδιου χρήστη. Για παράδειγμα, η χειρονομία "έλα εδώ" μπορεί να εκτελεστεί με μία κίνηση όλου του χεριού προς το χρήστη, με κίνηση μόνο των δαχτύλων, ή με κίνηση μόνο του δείκτη του χεριού (ενώ τα υπόλοιπα δάχτυλα είναι σε συστολή)· μπορεί να περιλαμβάνει χρήση και των δύο χεριών ή μόνο του ενός· μπορεί να προκύπτει και απλώς από ένα νεύμα του κεφαλιού. Μία χαρακτηριστική αναπαράσταση τέτοιων περιπτώσεων δίνεται στο σχήμα 1.1

Για όλους τους παραπάνω λόγους, η αυτόματη αναγνώριση χειρονομιών παραμένει ένα ιδιαίτερα ενεργό ερευνητικό πεδίο, με πολλές σχετικές δημοσιεύσεις [41, 86, 95], που δίνουν ώθηση στο state-of-the-art, και φιλοδοξούν να καινοτομήσουν σε ένα πρόβλημα με ερευνητικό, αλλά και πρακτικό ενδιαφέρον.



(α΄) (β΄) (γ΄)

(δ΄) (ε΄) (ϝ΄)

Σχήμα 1.1: Πάνω σειρά: Διαφορετικές εκτελέσεις του “έλα εδώ”, με κίνηση μόνο του δείκτη προς το χρήστη (α΄), με κίνηση όλων των δαχτύλων (β΄), και τέλος με κίνηση ολόκληρου του χεριού προς το χρήστη (γ΄)· Κάτω σειρά: Εκτέλεση του “έλα εδώ” με το δεξί χέρι (δ΄), με το αριστερό χέρι (ε΄), και με τα δύο ταυτόχρονα (ϝ΄).

1.2 Η σημασία της Πολυτροπικής Επεξεργασίας

Με κίνητρο τον τρόπο που οι άνθρωποι αντιλαμβάνονται το περιβάλλον τους, και ο οποίος βασίζεται σε επεξεργασία πολυτροπικής πληροφορίας [51], όλο και περισσότερες εφαρμογές εστιάζουν στην επεξεργασία διαφορετικών τροπικοτήτων, και στη σύμμειξη αυτών. Για παράδειγμα, η οπτική πληροφορία έχει χρησιμοποιηθεί με μεγάλη επιτυχία σε προβλήματα οπτικοακουστικού λόγου, βελτιώνοντας την αυτόματη αναγνώριση ομιλίας κάτω από δύσκολες ακουστικές συνθήκες [66]. Αντίστοιχα, η οπτικοακουστική παρακολούθηση του ανθρώπου είναι ιδιαίτερα χρήσιμη σε σενάρια παρακολούθησης ομιλητή που περιλαμβάνουν ανθρώπους που κινούνται χωρίς κάποιο περιορισμό σε περιβάλλοντα, όπως χώρους συνάντησης ή έξυπνα δωμάτια [6, 14, 65]. Τέλος, ειδικά και για το πρόβλημα της αναγνώρισης χειρονομιών, που αποτελεί το κύριο πεδίο έρευνας αυτής της εργασίας, η εφαρμογή διαφόρων σχημάτων σύμμειξης της οπτικής με την ηχητική πληροφορία, έχει αποδειχθεί ότι προσδίδει βελτιωμένα αποτελέσματα σε σύγκριση με την ανεξάρτητη επεξεργασία των δύο τροπικοτήτων [64].

Υπό την επίδραση των παραπάνω ερευνητικών επιτευγμάτων, τα πολυτροπικά συστήματα επικοινωνίας με χρήση χειρονομιών, έχουν προσελκύσει εντόνως την προσοχή της επιστημονικής κοινότητας τα τελευταία χρόνια [37, 84]. Αυτή η εξέλιξη αποδίδεται αφενός στην μεγάλη τεχνολογική πρόοδο που έχει σημειωθεί, όπως για παράδειγμα η ευρεία διάδοση των καμερών βάθους τύπου Kinect, και αφετέρου στην πρωτοποριακή έρευνα που έχει διεξαχθεί από την εποχή του "Put that there" του Bolt [9].

Όπως εύκολα μπορούμε να αντιληφθούμε, η φυσική αίσθηση της αλληλεπίδρασης με ένα σύστημα με τη χρήση χειρονομιών μπορεί να ενισχυθεί σημαντικά από την παρουσία πολλαπλών τροπικοτήτων. Οι στατικές και οι δυναμικές χειρονομίες, η μορφή του χεριού, καθώς και η ομιλία, συνθέτουν όλα μαζί ένα ελκυστικό σύνολο από ροές πληροφορίας, που εμφανίζουν σημαντικά πλεονεκτήματα για την αλληλεπίδραση ανθρώπου-μηχανής [62]. Όλα τα παραπάνω, θέτουν ένα σύνολο από ερευνητικές προκλήσεις για τον εντοπισμό της χρήσιμης πληροφορίας στα οπτικά και ακουστικά σήματα, την εξαγωγή των κατάλληλων χαρακτηριστικών, την εφαρμογή αποδοτικών ταξινομητών, αλλά και τον πολυτροπικό συνδυασμό των διαφορετικών πηγών πληροφορίας μέσω διαφόρων σχημάτων σύμμειξης [37].

1.3 Σχετική βιβλιογραφία

Σε αυτή την υποενότητα, παρουσιάζουμε ένα μέρος της σημαντικότερης βιβλιογραφίας που αφορά την αναγνώριση χειρονομιών, αλλά και συγγενή ερευνητικά προβλήματα. Πιο συγκεκριμένα, επιλέξαμε να κάνουμε μια

παρουσίαση σε τρία, εν γένει διαφορετικά, αλλά συγγενικά πεδία

1. Εμπνευσμένοι από τις εργασίες που η αναγνώριση χειρονομιών αντιμετωπίζεται σε ένα ενιαίο (ή παρόμοιο) πλαίσιο με την αναγνώριση νοηματικής γλώσσας, αρχικά παρουσιάζουμε ένα κομμάτι της σχετικής εργασίας που έχει γίνει από κοινού στα δύο προβλήματα. Ασφαλώς και υπάρχουν διαφορές στα δύο προβλήματα, μιας και η νοηματική γλώσσα είναι πολύ πιο πολύπλοκη και συγκεκριμένη από τις αδρές και "γενικότερου σκοπού" χειρονομίες, ωστόσο και στις δύο περιπτώσεις εστιάζουμε σε κινήσεις των άνω άκρων των χρηστών, οπότε μας επιτρέπεται να χρησιμοποιήσουμε (μέχρι ένα σημείο) κοινές μεθοδολογίες.

2. Παράλληλα, οι εργασίες που εστιάζουν τόσο σε αναγνώριση χειρονομιών, όσο και σε αναγνώριση ανθρωπίνων δράσεων, μας επιτρέπουν να παρουσιάσουμε την ερευνητική πρόοδο που σχετίζεται και με τα δύο αυτά πεδία. Τόσο οι χειρονομίες, όσο και οι ανθρώπινες δράσεις, αφορούν κινήσεις των ανθρώπων, με τη (σημαντική) διαφορά ότι στις χειρονομίες οι κινήσεις είναι εστιασμένες στα χέρια, ενώ στις γενικότερες δράσεις αφορούν διάφορα μέρη του σώματος. Ωστόσο, λόγω της γενικότητας των μεθόδων της αναγνώρισης δράσεων, είναι εφικτό να χρησιμοποιηθούν στο ειδικότερο πλαίσιο της αναγνώρισης χειρονομιών.

3. Τέλος, δίνουμε έμφαση σε βιβλιογραφικές αναφορές που σχετίζονται με τις μεθόδους πολυτροπικής σύμμειξης. Ασφαλώς κυριαρχεί το κομμάτι της πολυτροπικής αναγνώρισης χειρονομιών, όμως δεν περιοριζόμαστε εκεί, στοχεύοντας να παρουσιάσουμε την ευρύτερη φιλοσοφία ενός συστήματος πολυτροπικής σύμμειξης, αλλά και τις απαιτήσεις του.

1.3.1 Αναγνώριση Νοηματικής Γλώσσας και Χειρονομιών

Το πρόβλημα της αναγνώρισης νοηματικής γλώσσας είναι ένα ιδιαίτερα πολύπλοκο πρόβλημα, που περιλαμβάνει ένα πολύ ευρύ λεξιλόγιο, μια αυστηρά ορισμένη γραμματική, και διέπεται από πολύ συγκεκριμένες κινήσεις των χεριών, θέσεις, αλλά και χειρομορφές. Στην επόμενη υποενότητα θα αποφύγουμε να ασχοληθούμε με τις ιδιαίτερες δυσκολίες που παρουσιάζει και την ειδική διαχείριση που απαιτεί. Αντ' αυτού θα εστιάσουμε στη μεθοδολογία εργασίας στο σημείο που αυτή είναι κοινή με την αναγνώριση χειρονομιών, δηλαδή κυρίως στο σύστημα επεξεργασίας και ανάλυσης των κινήσεων των χεριών.

Το πρώτο σημαντικό βήμα ενός τέτοιου συστήματος, είναι ο εντοπισμός των χεριών. Αυτό υλοποιείται συνήθως με τη χρήση διαφόρων οπτικών χαρακτηριστικών, όπως το χρώμα του δέρματος, οι ακμές, το σχήμα, η κίνηση, αλλά και με συνδυασμούς αυτών. Η χρωματική πληροφορία αποδεικνύεται συχνά

χρήσιμη, λόγω του χαρακτηριστικού χρώματος του ανθρώπινου δέρματος, και χρησιμοποιείται σε πολλές περιπτώσεις για την κατάτμηση των περιοχών δέρματος, και τον εντοπισμό των χεριών [3, 99, 75]. Όσον αφορά την ευρωστία στις αλλαγές της φωτεινότητας, αυτή επιτυγχάνεται με την επιλογή χρωματικών χώρων όπως ο *HSV*, ο *YCbCr* ή ο *CIE-Lab*, που διαχωρίζουν τη συνιστώσα του χρώματος από αυτή της φωτεινότητας [82, 38].

Το επόμενο βήμα της οπτικής επεξεργασίας είναι η παρακολούθηση των χεριών, η οποία βασίζεται συνήθως σε blobs [79, 81, 3], στην εμφάνιση του χεριού [35], ή στο περίγραμμά του [13, 20]. Να σημειώσουμε ωστόσο, ότι κάποιες πιο πρόσφατες δουλειές ξεπερνούν τα προβλήματα και του εντοπισμού, αλλά και της παρακολούθησης του χεριού, εκμεταλλευόμενες την παρακολούθηση του ανθρώπινου σκελετού που παρέχει ένας αισθητήρας τύπου Kinect [69].

Επιπλέον, ένα κρίσιμο ζήτημα που πρέπει να αντιμετωπίσει τόσο ένα σύστημα αναγνώρισης νοηματικής γλώσσας, όσο και ένα σύστημα αναγνώρισης χειρονομιών, είναι η εξαγωγή χαρακτηριστικών από την περιοχή του χεριού. Ένα πολύ συνηθισμένο χαρακτηριστικό είναι το δισδιάστατο (2D) ή τρισδιάστατο (3D) κέντρο βάρους της περιοχής του χεριού [79, 4, 81, 20], καθώς και χαρακτηριστικά κίνησης [99, 13]. Αρκετές δουλειές χρησιμοποιούν επίσης γεωμετρικά χαρακτηριστικά που σχετίζονται με το χέρι, όπως οι ροπές σχήματος [34, 79], τα μεγέθη και οι αποστάσεις μεταξύ των δαχτύλων, της παλάμης, καθώς και της πίσω περιοχής των χεριών [4]. Σε άλλες περιπτώσεις, το περίγραμμα του χεριού χρησιμοποιείται για την εξαγωγή χαρακτηριστικών που παραμένουν αναλλοίωτα σε μετασχηματισμούς παράλληλης μετατόπισης, αλλαγής κλίμακας και περιστροφής, όπως επί παραδείγματι οι περιγραφητές Fourier [13, 18].

Οι εικόνες των χεριών μετά την κατάτμηση, συνήθως κανονικοποιούνται ως προς το μέγεθος, τον προσανατολισμό στο χώρο, και/ή την φωτεινότητα, ενώ στη συνέχεια εφαρμόζεται PCA (Principal Component Analysis) με σκοπό την μείωση της διαστασιμότητας και την αναπαράσταση της χειρομορφής μέσω περιγραφητών [7, 20, 96, 24, 26]. Συσχετιζόμενη με τις προσεγγίσεις με PCA είναι και η χρήση ενεργών μοντέλων σχήματος και εμφάνισης [19, 50], για την εξαγωγή χαρακτηριστικών επί της χειρομορφής [2, 35, 11, 29, 71]

Όσον αφορά τώρα το θέμα της στατιστικής μοντελοποίησης, τα κρυφά Μαρκοβιανά μοντέλα (HMMs) αποτελούν ίσως το πιο δημοφιλές αλλά και το πιο ισχυρό εργαλείο, και έχουν χρησιμοποιηθεί με επιτυχία τόσο για την αναγνώριση χειρονομιών [53, 54, 83, 97] όσο και για την αναγνώριση νοηματικής γλώσσας [78, 59]. Μάλιστα, πολλές ενδιαφέρουσες εφαρμογές και επεκτάσεις έχουν αναπτυχθεί με βάση τα HMMs. Για παράδειγμα, έχει περιγραφεί ένα μοντέλο κατωφλίου [45], με σκοπό τον εντοπισμό χειρονομιών και την ορθή αντιμετώπιση δεδομένων που δεν περιλαμβάνουν χειρονομίες, ενώ έχει εισαχθεί και η έννοια των παραμετρικών HMMs [94], για παραμετρικές χειρονομίες (με συστηματικές

δηλαδή χωρικές παραλλαγές). Τέλος, μία ακόμα σημαντική συνεισφορά είναι αυτή των παράλληλων HMMs (Parallel HMMs ή PaHMMs) [85], τα οποία αντιμετωπίζουν την ανάγκη για χρήση πολλών ροών πληροφορίας ταυτόχρονα, και παρέχουν ένα αποδοτικό σχήμα σύμμειξης.

1.3.2 Αναγνώριση Δράσεων και Χειρονομιών

Το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων σε βίντεο έχει συγκεντρώσει πολύ μεγάλο ερευνητικό ενδιαφέρον τα τελευταία χρόνια, και ειδικά την τελευταία δεκαετία έχουν σημειωθεί εντυπωσιακά βήματα προόδου. Πιο συγκεκριμένα, με σκοπό την αναγνώριση ρεαλιστικών δράσεων σε μη ελεγχόμενα περιβάλλοντα, έχουν αναπτυχθεί αρκετά γενικές μέθοδοι οι οποίες θέτουν ελάχιστους περιορισμούς επί των δεδομένων εισόδου, και μπορούν να χρησιμοποιηθούν με την ίδια ευκολία και σε δεδομένα αναγνώρισης χειρονομιών.

Η πλέον δημοφιλής μεθοδολογία σε αυτή την περιοχή είναι αυτή των χωροχρονικών σημείων ενδιαφέροντος (Spatio-Temporal Interest Points ή STIP), που προτάθηκε από το Laptev [42], και περιλαμβάνει τέσσερα βήματα:

1. Εντοπισμός χωροχρονικών σημείων ενδιαφέροντος στην ακολουθία βίντεο.

2. Υπολογισμός περιγραφητών στις γειτονιές των σημείων που ανιχνεύθηκαν στο προηγούμενο βήμα.

3. Αναπαράσταση της ακολουθίας βίντεο σε ένα ιστόγραμμα "οπτικών λέξεων" με χρήση της τεχνικής Bag-of-Features (εμπνευσμένη από τη μέθοδο Bag-of-Words του πεδίου της επεξεργασίας κειμένου).

4. Ταξινόμηση της ακολουθίας βίντεο σε κάποια κατηγορία δράσεων (για παράδειγμα με χρήση του ταξινομητή SVM).

Στα χρόνια που ακολούθησαν τη δημοσίευση της παραπάνω εργασίας του Laptev, παρουσιάστηκαν αρκετές ακόμα εργασίες βασισμένες στην ίδια μεθοδολογία.

Κατ' αρχάς, όσον αφορά τον εντοπισμό των χωροχρονικών σημείων ενδιαφέροντος, ο ίδιος ο Laptev είχε προτείνει μία επέκταση του ανιχνευτή Harris στον τρισδιάστατο χώρο, τον Harris3D, για την ανίχνευση χωροχρονικών γωνιών. Άλλες μέθοδοι βασίζονται στη χρήση φίλτρων Gabor, όπως ο ανιχνευτής Cuboid [23], ή ο Gabor3D [47]. Τέλος έχει προταθεί και ο ανιχνευτής Hessian [92], οι ανιχνεύσεις του οποίου αντιστοιχούν στα τοπικά μέγιστα της ορίζουσας της τρισδιάστατης χωροχρονικής μήτρας Hessian.

Αντίστοιχα με τις προσπάθειες σε ανιχνευτές σημείων ενδιαφέροντος, στη βιβλιογραφία υπάρχει εκτεταμένη έρευνα και ανάλυση σχετικά με το είδος των περιγραφητών που θα χρησιμοποιηθούν στις γειτονιές αυτών των σημείων.

Ένα μεγάλο κομμάτι αφορά τη χρήση δημοφιλών περιγραφητών βασισμένων σε gradients, (HOG), ή στην οπτική ροή (HOF), καθώς και συνδυασμού αυτών των επιμέρους περιγραφητών, HOG/HOF [43]. Επιπλέον, έχουν γίνει αρκετές προσπάθειες για την επέκταση δισδιάστατων περιγραφητών στις τρεις διαστάσεις, όπως για παράδειγμα ο 3D-SIFT [74], ο HOG3D [39], καθώς και η επέκταση του SURF περιγραφητή, ο E-SURF [92].

Πέρα από τα παραπάνω, κάποιες πιο πρόσφατες εργασίες προτείνουν αλλαγές ή επεκτάσεις της συνηθισμένης μεθοδολογίας των χωροχρονικών σημείων ενδιαφέροντος. Κατ' αρχάς, έχει προταθεί η χρήση πυκνής δειγματοληψίας για την εξαγωγή χαρακτηριστικών, και σύμφωνα με εκτενή μελέτη [89], φαίνεται να υπερέχει έναντι της ανίχνευσης χωροχρονικών σημείων ενδιαφέροντος. Επιπλέον, λόγω των διαφορετικών χαρακτηριστικών που εμφανίζει το δισδιάστατο πεδίο της εικόνας, από το μονοδιάστατο πεδίο του χρόνου, έχει προταθεί η παρακολούθηση σημείων ενδιαφέροντος στις ακολουθίες βίντεο, αντί για εντοπισμό τέτοιων στον κοινό 3D χώρο. Αυτή η παρακολούθηση μπορεί να γίνει είτε με χρήση του αλγορίθμου Lucas-Kanade [52], είτε με αντιστοίχηση περιγραφητών SIFT σε διαδοχικά καρέ του βίντεο [80], είτε ακόμα και με εξαγωγή πυκνών τροχιών [87]. Τέλος, ιδιαίτερο ενδιαφέρον παρουσιάζει και η προσπάθεια για την εξάλειψη της κίνησης της κάμερας [88], η οποία είναι ιδιαίτερα συνηθισμένη σε βάσεις ρεαλιστικών δεδομένων.

Ακόμα και αν οι παραπάνω ιδέες και μεθοδολογίες εργασίας είναι αρκετά γενικές ώστε να εφαρμοστούν και σε προβλήματα αναγνώρισης χειρονομιών, οι παραθέσεις που δίνουμε ασχολούνται σχεδόν αποκλειστικά με την αναγνώριση δράσεων. Ωστόσο, υπάρχουν και κάποιες πιο πρόσφατες εργασίες, οι οποίες αντιμετωπίζουν την αναγνώριση δράσεων και χειρονομιών σε ένα ενιαίο πλαίσιο. Μάλιστα, ο κύριος όγκος της δουλειάς, σε αντίθεση με ότι παρουσιάσαμε μέχρι τώρα, έχει γίνει σε δεδομένα βάθους προερχόμενα από κατάλληλες κάμερες όπως το Kinect.

Μία από τις πρώτες εργασίες ήταν αυτή των Li et al. [46], οι οποίοι προτείνουν τον υπολογισμό ενός συνόλου 3D σημείων (σε αναλογία με τη μέθοδο του Bag-of-Words). Για να το πετύχουν, δειγματοληπτούν σημεία από τη σιλουέτα της εικόνας βάθους, και στη συνέχεια ομαδοποιούν αυτά τα σημεία σε clusters, ώστε να προκύψουν κάποιες σημαντικές/κύριες πόζες, που αποτελούν και το λεξιλόγιό τους. Από την άλλη, στην εργασία των Wang et al. [90] επιλέγεται ένα πλήθος τρισδιάστατων χωροχρονικών όγκων από όλους τους διαθέσιμους χωροχρονικούς όγκους των ακολουθιών δεδομένων βάθους. Η επιλογή αυτών των χωροχρονικών τμημάτων γίνεται με τη χρήση LDA, ώστε να διατηρηθούν οι πλέον διαχωρίσιμοι από αυτούς.

Μια διαφορετική προσέγγιση παρουσιάζουν οι Yang et al., οι οποίοι στην εργασία τους [101] ακολουθούν μια πιο ολιστική μέθοδο, με την οποία αντί να

χρησιμοποιούν τοπικά σημεία ενδιαφέροντος, εξάγουν έναν ολικό περιγραφητή για κάθε ακολουθία. Πιο συγκεκριμένα μία ακολουθία βίντεο συνοψίζεται σε μία μόνο εικόνα, τον χάρτη κίνησης, ο οποίος αντιπροσωπεύει τη μέση διαφορά μεταξύ των καρέ της ακολουθίας βάθους. Εν συνεχεία, ένας απλός HOG περιγραφητής υπολογίζεται από αυτό το χάρτη.

Παράλληλα, η αναγνώριση χειρονομιών και γενικότερα ανθρωπίνων δράσεων έχει βοηθηθεί ιδιαίτερα από την ακριβή και αποδοτική εκτίμηση της ανθρώπινης πόζας, με χρήση δεδομένων βάθους [76], ένα πρόβλημα ιδιαίτερα απαιτητικό όταν χρησιμοποιούνται αποκλειστικά RGB ακολουθίες εικόνων. Εστιάζοντας στο σκελετό, οι Wang et al. [91] χρησιμοποιούν τις αρθρώσεις ως σημεία ενδιαφέροντος. Πιο συγκεκριμένα τα χαρακτηριστικά που εξάγουν περιλαμβάνουν τόσο πληροφορία για το σχήμα της περιοχής που περιβάλλει την άρθρωση, όσο και πληροφορία για την ίδια την τοποθεσία της άρθρωσης. Αυτά τα χαρακτηριστικά υπολογίζονται σε κάθε καρέ του βίντεο, και οι συντελεστές του μετασχηματισμού Fourier χρησιμοποιούνται για να περιγράψουν την χρονική μεταβολή τους. Επίσης σε μια άλλη εργασία των Yang και Tian [100], χρησιμοποιείται μια μειωμένης διαστασιμότητας αναπαράσταση της πληροφορίας των σημείων του σκελετού, που προκύπτει από τις χωροχρονικές αποστάσεις των εν λόγω σημείων και συνδυάζει πληροφορία που αφορά τόσο τη στατική πόζα, όσο και την κίνηση.

Τέλος, η πιο πρόσφατη δουλειά στην ευρύτερη περιοχή είναι αυτή των Oreifej και Liu [60], οι οποίοι προσπαθούν να συμπεριλάβουν πληροφορία τόσο από το κανάλι του σχήματος, όσο και από αυτό της κίνησης σε ένα ιστόγραμμα, το οποίο υπολογίζεται από τον τετραδιάστατο χώρο των χωρικών συντεταγμένων, του βάθους, και του χρόνου. Η συγκεκριμένη εργασία μάλιστα, παρουσιάζει state-of-the-art αποτελέσματα για όλες τις σχετικές βάσεις χειρονομιών και δράσεων σε δεδομένα βάθους.

1.3.3 Πολυτροπική Αναγνώριση Χειρονομιών

Η πολυτροπική αναγνώριση χειρονομιών, και γενικότερα τα πολυτροπικά συστήματα επικοινωνίας, έχουν συγκεντρώσει ιδιαίτερο ενδιαφέρον στη βιβλιογραφία [37, 84], μιας και η ενίσχυση της παρεχόμενης πληροφορίας με περισσότερες από μία τροπικότητες μπορεί να διευκολύνει διαισθητικά ένα πρόβλημα. Για παράδειγμα, όταν ένα σύστημα επικοινωνίας ανθρώπου υπολογιστή μέσω χειρονομιών εμπλουτιστεί με την τροπικότητα του ήχου, μειώνεται η δυσκολία του προβλήματος και αυξάνει η φυσικότητα του συστήματος [62]. Ωστόσο, η ερευνητική πρόκληση στην ανάπτυξη ενός τέτοιου συστήματος παραμένει, διότι απαιτείται πολύ καλός χειρισμός όλων των διαφορετικών τροπικοτήτων, καθώς και η σύμμειξη αυτών. Και πέρα από τα τεχνικά προβλήματα που πρέπει να επιλυθούν, αυτό που έχει το μεγαλύτερο ερευνητικό ενδιαφέρον είναι το σχήμα της πολυτροπικής σύμμειξης· *πως* και *πότε* θα πραγματοποιηθεί η σύμμειξη στην

επεξεργασία του συστήματος. Με βάση την απάντηση σε αυτές τις ερωτήσεις, μια συνήθης κατηγοριοποίηση των μεθόδων σύμμειξης είναι στις επόμενες τρεις κλάσεις:

- Σε μεθόδους πρώιμης σύμμειξης (early fusion), ή σύμμειξης επιπέδου χαρακτηριστικών (feature-level fusion). Σε αυτές τις περιπτώσεις υπάρχει συνδυασμός των δεδομένων από τις διάφορες τροπικότητες, με αποτέλεσμα να προκύπτει ένα μεικτό διάνυσμα χαρακτηριστικών, με βάση το οποίο λαμβάνεται η πολυτροπική απόφαση. Ένα παράδειγμα αποτελεί η συνένωση χαρακτηριστικών εικόνας και ήχου που επιχειρούν οι Adjoudani και Benoit [1].

- Σε μεθόδους σύμμειξης στο τελικό στάδιο (late fusion), ή σύμμειξης επιπέδου απόφασης (decision-level fusion). Σε αυτές τις περιπτώσεις κάθε τροπικότητα επεξεργάζεται ξεχωριστά και λαμβάνεται μια απόφαση ανεξάρτητα από τις υπόλοιπες τροπικότητες. Το σχήμα σύμμειξης τότε έχει σκοπό να αξιολογήσει τις επιμέρους αποφάσεις και να τις συνδυάσει για την εξαγωγή της τελικής πολυτροπικής απόφασης. Ένα παράδειγμα αποτελούν τα PaHMMs των Vogler και Metaxas [85] για τη σύμμειξη διαφορετικών ροών πληροφορίας κατά την αναγνώριση νοηματικής γλώσσας.

- Σε μεθόδους ενδιάμεσης σύμμειξης (intermediate ή mid-level fusion), οι οποίες αποτελούν ένα συμβιβασμό των δύο προηγούμενων περιπτώσεων, όπου και επιτρέπεται ένας βαθμός επεξεργασίας της κάθε τροπικότητας (ή και πιθανώς ταξινόμησης με βάση αυτή), πριν πραγματοποιηθεί η τελική σύμμειξη όλων των τροπικοτήτων. Παράδειγμα αποτελεί η εργασία των Cohen et al, για την αναγνώριση εκφράσεων του προσώπου με χρήση ιεραρχικών σχημάτων HMMs [17].

Παρά τις δυσκολίες και τα διλήμματα που θα πρέπει να αντιμετωπιστούν, έχουν γίνει σημαντικά βήματα στην κατεύθυνση πολυτροπικών συστημάτων επικοινωνίας, ξεκινώντας ήδη από την αρχή της δεκαετίας του '80. Πρωτοποριακή και θεμελιώδης θεωρείται η εργασία του Bolt [9], που παρουσιάζει ένα δωμάτιο πολυτροπικής επικοινωνίας, όπου ο χρήστης προέβαινε σε ταυτόχρονη χρήση λόγου και χειρονομιών για τη διατύπωση απλών εντολών (όπως το διάσημο "Put-that-there"). Το ενδιαφέρον ήταν ότι καμία εντολή δεν μπορούσε να ερμηνευτεί και να γίνει πλήρως κατανοητή χωρίς τη χρήση και των δύο τροπικοτήτων, κάνοντας την πολυτροπική σύμμειξη ζωτικό συστατικό της λειτουργίας του συστήματος. Την εργασία του Bolt ακολούθησαν και άλλες δημοσιεύσεις που περιέγραφαν συστήματα πολυτροπικής επικοινωνίας, όπως αυτό των Neal et al. [56] που επέτρεπε την επικοινωνία με χρήση προφορικής ή γραπτής φυσικής γλώσσας και

χειρονομιών, αλλά και αυτό των Koons et al. [40] που συνδύαζε την επεξεργασία ομιλίας, χειρονομιών και βλέμματος (eye gaze).

Αυτές οι προσπάθειες έχουν αποκτήσει νέο ενδιαφέρον, με την εισαγωγή των αισθητήρων 3D όρασης, όπως το Kinect, μιας και οι ευκαιρίες και οι δυνατότητες για ενασχόληση με ερευνητικά προβλήματα στο πεδίο της πολυτροπικής σύμμειξης, είναι ακόμα περισσότερες. Λαμβάνοντας υπόψη και την έρευνα στα πλαίσια της παρούσας διπλωματικής, ιδιαίτερο ενδιαφέρον παρουσιάζουν εργασίες σε βάσεις που περιλαμβάνουν οπτικά δεδομένα εκτέλεσης των χειρονομιών, συνοδευόμενα από τις αντίστοιχες φωνητικές εντολές/εκφράσεις.

Σε αυτή την κατεύθυνση, οι Wu et al. [95] καθοδηγούνται κυρίως από την ακουστική πληροφορία, την οποία χρησιμοποιούν για να εντοπίσουν τα όρια της ηχητικής εντολής, και κατ' επέκταση της χειρονομίας. Στη συνέχεια, χρησιμοποιούν ανεξάρτητα δύο ταξινομητές, έναν με χρήση πληροφορίας από τα χαρακτηριστικά του σκελετού, και ένα δεύτερο που εκμεταλλεύεται την ηχητική πληροφορία. Τέλος, γίνεται ο συνδυασμός των σκορ των δύο ταξινομητών για τη λήψη της τελικής απόφασης σε κάθε τμήμα.

Παρόμοια είναι και η μέθοδος των Bayer και Silbermann [5], οι οποίοι χρησιμοποιούν την ακουστική πληροφορία για τον εντοπισμό των ηχητικών εντολών, και κάνουν χρήση δύο διαφορετικών ταξινομητών με χαρακτηριστικά προερχόμενα από τον ήχο. Για την οπτική πληροφορία από την άλλη, χρησιμοποιούν κυλιόμενα παράθυρα για να υπολογίσουν την πιθανότητα κάθε καρέ της ακολουθίας να ανήκει σε κάποια κατηγορία χειρονομιών. Τέλος προκύπτει ένα σταθμισμένο άθροισμα των πιθανοτήτων των διαφόρων ταξινομητών (δύο για τον ήχο, ένα για το σκελετό) για τη λήψη απόφασης.

Οι Nandakumar et al [55] αρχικά συνδυάζουν την πληροφορία από το σκελετό του χρήστη και την ηχητική πληροφορία για τη χρονική κατάτμηση της ακολουθίας των χειρονομιών. Στη συνέχεια εφαρμόζουν τρεις ταξινομητές, χρησιμοποιώντας αντίστοιχα την ηχητική πληροφορία, την οπτική πληροφορία από το RGB βίντεο, και την οπτική πληροφορία από το σκελετό του χρήστη. Η αναγνώριση της ακολουθίας χειρονομιών προκύπτει από τη στάθμιση των αποφάσεων των τριών ταξινομητών.

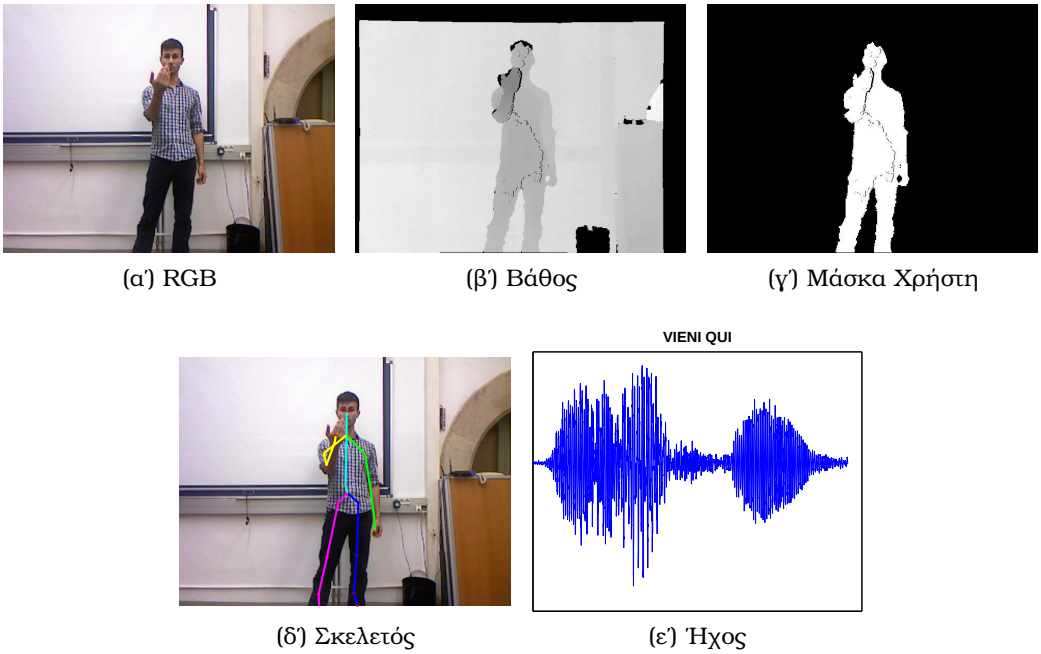
Αξίζει να σημειωθεί ότι και οι τρεις ανωτέρω μέθοδοι έχουν εφαρμοστεί στη βάση πολυτροπικών χειρονομιών ChaLearn, που παρουσιάζουμε στην υποενότητα 1.5.2 και την οποία έχουμε χρησιμοποιήσει εκτενώς στους πειραματισμούς μας.

1.4 Ο αισθητήρας Kinect

Ο αισθητήρας Microsoft Kinect αποτελεί μία κάμερα, που επιτρέπει τη λήψη δεδομένων βάθους, εκτός από συνηθισμένα RGB δεδομένα, δίνοντας έτσι την ευκαιρία για μια τρισδιάστατη αναπαράσταση του χώρου που καταγράφεται. Το

Kinect δεν είναι η μοναδική κάμερα που έχει τη δυνατότητα να καταγράψει το βάθος της σκηνής (άλλα παραδείγματα αποτελούν οι αισθητήρες Xtion PRO και Bumblebee), ωστόσο λόγω του χαμηλού κόστους της και της ευρείας διάδοσής της αποτελεί το επίκεντρο πολλών ερευνητικών προσπαθειών ιδιαίτερα στα πεδία της όρασης υπολογιστών [32], της ρομποτικής, αλλά και της επικοινωνίας ανθρώπου-μηχανής.

Η καινοτομία βέβαια του αισθητήρα Kinect δεν περιορίζεται στη 3D απεικόνιση του χώρου λήψης. Με χρήση ενσωματωμένου λογισμικού και καινοτόμων αλγορίθμων [76], καθιστά δυνατή την παρακολούθηση του σκελετού του χρήστη που βρίσκεται στο πεδίο λήψης του. Ο σκελετός αυτός αποτελείται από 20 σημεία, προκύπτει από το χάρτη βάθους, και στην ουσία αντιμετωπίζει το πρόβλημα της εκτίμησης πόζας του ανθρώπου, ένα πρόβλημα ιδιαίτερα απαιτητικό με χρήση συμβατικών RGB εικόνων [102, 72]. Παράλληλα, πάλι με χρήση του χάρτη βάθους, εντοπίζει το περίγραμμα των ανθρώπων που βρίσκονται εντός του πεδίου λήψης, ενώ χάρη στη συστοιχία μικροφώνων που είναι ενσωματωμένη στη συσκευή, προσφέρει και παράλληλη καταγραφή ηχητικών δεδομένων. Μία πλήρης απεικόνιση των πλούσιων δεδομένων που είναι σε θέση να καταγράψει, δίνεται στο σχήμα 1.2.



(α') RGB (β') Βάθος (γ') Μάσκα Χρήστη

(δ') Σκελετός (ε') Ήχος

Σχήμα 1.2: Πλήρης λίστα των ροών πληροφορίας που έχει τη δυνατότητα να καταγράψει ο αισθητήρας Kinect. Τα δεδομένα προέρχονται από στιγμιότυπο της πολυτροπικής βάσης χειρονομιών ChaLearn [27].

Στο πεδίο της αναγνώρισης χειρονομιών, που αποτελεί και το ερευνητικό επίκεντρο της παρούσας διπλωματικής, το Kinect έχει χρησιμοποιηθεί σε πολλές ερευνητικές προσπάθειες [41], κυρίως λόγω της καταγραφής δεδομένων βάθους και της επιστροφής του σκελετού του χρήστη σε πραγματικό χρόνο. Ωστόσο, η αναγνώριση χειρονομιών δεν είναι το μόνο ερευνητικό πρόβλημα για το οποίο υπάρχει συνεχιζόμενη έρευνα με χρήση του αισθητήρα Kinect. Έχει χρησιμοποιηθεί επίσης σε προβλήματα:

- εντοπισμού και παρακολούθησης αντικειμένων και ανθρώπων [98].
- αναγνώρισης αντικειμένων [8] και κατηγοριοποίησης σκηνής [77].
- εκτίμησης πόζας ανθρώπου [30].
- αναγνώρισης δράσεων [46].
- εντοπισμού και εκτίμησης πόζας χεριών [58].
- τρισδιάστατης χαρτογράφησης εσωτερικών χώρων [57].

Στη συνέχεια της διπλωματικής, θα εκμεταλλευτούμε σε πολλά σημεία τις ειδικές δυνατότητες του αισθητήρα Kinect, είτε πρόκειται για το βάθος, είτε για το σκελετό, είτε για την παράλληλη καταγραφή ηχητικής και οπτικής πληροφορίας. Μάλιστα τα δεδομένα και από τις τρεις βάσεις δεδομένων που εργαστήκαμε έχουν προέλθει αποκλειστικά από αισθητήρες αυτού του είδους, γι' αυτό άλλωστε θεωρήσαμε και σκόπιμο να κάνουμε μια σύντομη παρουσίασή του σε αυτό το σημείο.

1.5 Διαθέσιμες βάσεις δεδομένων

Η έρευνα στο ευρύτερο πεδίο της όρασης υπολογιστών έχει συνεισφέρει στην παρουσία πολλών διαθέσιμων βάσεων δεδομένων για τον απαιτούμενο πειραματισμό στα διάφορα ερευνητικά προβλήματα. Για παράδειγμα, αρκετές βάσεις-benchmark είναι διαθέσιμες για την αναγνώριση ανθρώπινων δράσεων [49, 73], αλλά και για την αναγνώριση νοηματικής γλώσσας [25]. Αντίστοιχα, στο πεδίο της αναγνώρισης χειρονομιών, πολλές βάσεις δεδομένων έχουν γίνει διαθέσιμες, με κάποιες από τις πιο πρόσφατες να έχουν βιντεοσκοπηθεί με τη χρήση Kinect [31, 41], κάνοντας διαθέσιμη και την πληροφορία του βάθους. Παρ' όλα αυτά, λίγες είναι οι προσπάθειες που έχουν γίνει για καταγραφή πολυτροπικών δεδομένων, και δη για το πεδίο της αναγνώρισης χειρονομιών. Στη συνέχεια παρουσιάζουμε δύο τέτοιες βάσεις οι οποίες αποτέλεσαν και το κύριο ενδιαφέρον της έρευνάς μας στα πλαίσια αυτής της διπλωματικής, μαζί με μία τρίτη, η οποία χρησιμοποιήθηκε για προκαταρκτικά κυρίως πειράματα.

1.5.1 Η βάση στατικών χειρομορφών

Πρώτα θα παρουσιάσουμε μια βάση δεδομένων στατικών χειρομορφών, στην οποία δοκιμάστηκαν κάποια προκαταρκτικά πειράματα στα πλαίσια αυτής της διπλωματικής. Σε κάθε περίπτωση, η ταξινόμηση στατικών χειρομορφών είναι ένα διαφορετικής φύσης πρόβλημα από αυτό της αναγνώρισης χειρονομιών (μιας και δεν εμφανίζεται η διάσταση του χρόνου), ωστόσο έχει ενδιαφέρον να συγκρίνουμε την απόδοση κάποιων μεθόδων εξαγωγής χαρακτηριστικών και σε ένα διαφορετικό πλαίσιο.

Λεξιλόγιο

Ως λεξιλόγιο της συγκεκριμένης βάσης εστιάσαμε στις χειρομορφές που χρησιμοποιούνται στην ελληνική νοηματική γλώσσα για την αναπαράσταση των γραμμάτων του ελληνικού αλφαβήτου. Συνολικά, έχουμε διαθέσιμες 24 χειρομορφές οι οποίες παρουσιάζονται στο σχήμα 1.3 χρησιμοποιώντας για την απεικόνιση στιγμιότυπα της εν λόγω βάσης.

Δομή των Δεδομένων

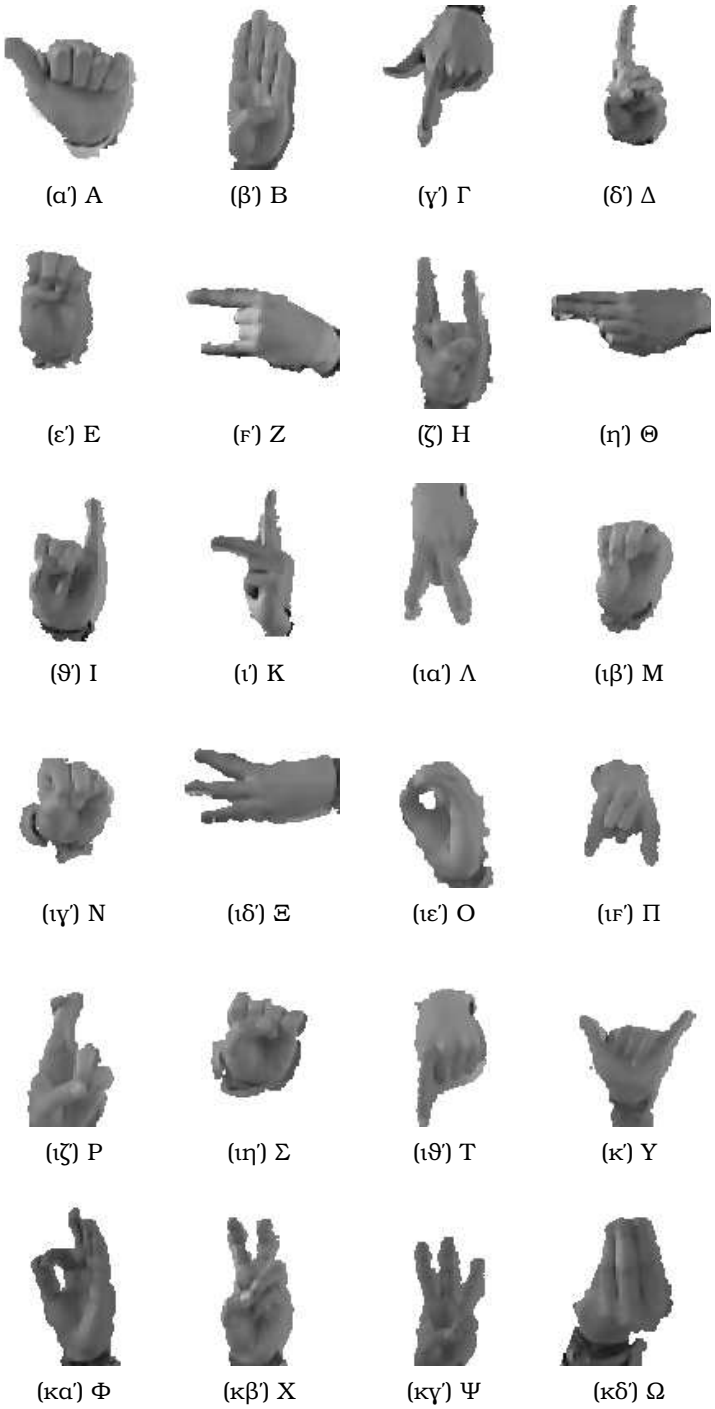
Συνολικά υπάρχουν διαθέσιμα δεδομένα από τέσσερις χρήστες (τρεις άνδρες και μία γυναίκα). Για κάθε χειρομορφή υπάρχουν τουλάχιστον 100 επαναλήψεις (αθροιστικά, και από τους τέσσερις χρήστες), ενώ ο συνολικός αριθμός των διαθέσιμων καρέ φτάνει στα 5202. Η λήψη έγινε με χρήση του αισθητήρα Kinect και τα δεδομένα περιλαμβάνουν μόνο γκρίζες εικόνες. Χρησιμοποιήθηκε ο σκελετός που επιστρέφει το Kinect για τον εντοπισμό της περιοχής του χεριού, ενώ με τη βοήθεια της πληροφορίας του βάθους έγινε κατάτμηση της χειρομορφής σε σχέση με το υπόλοιπο σώμα αλλά και το παρασκήνιο.

Πειραματισμοί

Για τους πειραματισμούς που θα παρουσιαστούν στη συνέχεια με χρήση της συγκεκριμένης βάσης χρησιμοποιήθηκε ένα μέρος των δεδομένων, που αφορούσε τις χειρομορφές για τα γράμματα Α έως Κ.

1.5.2 Η πολυτροπική βάση χειρονομιών ChaLearn

Εστιάζοντας στο κυρίως πρόβλημα της Αναγνώρισης Χειρονομιών, για τις ανάγκες των πειραμάτων, χρειαστήκαμε μία βάση δεδομένων, με εκτεταμένο λεξιλόγιο (περισσότερες από 10 χειρονομίες), πληθώρα διαφορετικών χρηστών, καθώς και αρκετές διαφορετικές επαναλήψεις που να αφορούν αυτές τις χειρονομίες. Με δεδομένες αυτές τις προδιαγραφές, χρησιμοποιήσαμε για τα πειράματά μας



Σχήμα 1.3: Στιγμιότυπα της βάσης δεδομένων στατικών χειρομορφών. Κάθε χειρομορφή αντιστοιχεί σε ένα γράμμα της ελληνικής αλφαβήτου, έτσι όπως αυτά εκφράζονται στην ελληνική νοηματική γλώσσα.

τη βάση δεδομένων του Διαγωνισμού Πολυτροπικής Αναγνώρισης Χειρονομιών "Multi-Modal Gesture Challenge 2013" που διοργανώθηκε από τον οργανισμό ChaLearn, που ειδικεύεται σε διαγωνισμούς στον ευρύτερο χώρο της Μηχανικής Μάθησης (Machine Learning). Η συγκεκριμένη βάση δεδομένων έχει καταγραφεί με χρήση του αισθητήρα Kinect και περιλαμβάνει ένα εκτεταμένο λεξιλόγιο 20 διαφορετικών χειρονομιών, οι οποίες εκτελούνται από 39 διαφορετικούς χρήστες, σε ένα φάσμα από 5 έως και 50 επαναλήψεις ανά χειρονομία και ανά χρήστη.

Παρεχόμενες τροπικότητες (modalities)

Ένα από τα πιο ενδιαφέροντα χαρακτηριστικά της συγκεκριμένης βάσης είναι η πολυτροπική (multimodal) φύση των δεδομένων. Χρησιμοποιώντας τις καινοτομικές λειτουργίες που συνδυάζει ο Kinect Sensor, έχει γίνει καταγραφή των διαφορετικών νοημάτων με ένα πλήθος διαφορετικών ροών πληροφορίας:

- Οπτική καταγραφή με χρήση RGB
- Οπτική καταγραφή με χρήση του χάρτη βάθους (Depth Image).
- Ηχητική καταγραφή από τη συστοιχία μικροφώνων.
- Πληροφορία εντοπισμού του χρήστη (User Index).
- Πληροφορία ανίχνευσης του σκελετού του χρήστη (Skeleton Tracking).

Συνεπώς υπάρχει μια πολύ πλούσια πληροφορία διαθέσιμη που ευνοεί και παροτρύνει τη σύμμειξη των διάφορων τροπικοτήτων.

Λεξιλόγιο

Όσον αφορά το λεξιλόγιο της βάσης μας, αυτό αποτελείται από 20 διαφορετικά νοήματα. Κάθε ένα από αυτά τα νοήματα αποτελεί συνδυασμό λεκτικής εκφοράς, καθώς και εκτέλεσης της αντίστοιχης χειρονομίας, και αντιστοιχεί σε 20 διαφορετικά ανθρωπολογικά-πολιτιστικά νοήματα της Ιταλικής Γλώσσας. Πιο συγκεκριμένα περιλαμβάνονται οι 20 παρακάτω χειρονομίες (αναφέρουμε την ιταλική φράση του κάθε νοήματος, μαζί με την ελληνική μετάφραση, αλλά και τη συντομογραφία του, έτσι όπως δηλαδή αναφέρεται το κάθε νόημα στη συνέχεια της διπλωματικής):

1. Vattene - Φύγε ("vattene")
2. Vieni Qui - Έλα εδώ ("vieniqui")
3. Perfetto! - Τέλεια! ("perfetto")

4. E' un furbo! - Είναι πονηρός ("furbo")
5. Che due palle - Είναι πολύ εκνευριστικό ("cheduepalle")
6. Che vuoi - Τι θες; ("chevuoi")
7. Vanno d'accordo - Τα πάνε καλά μεταξύ τους ("daccordo")
8. Sei pazzo - Είσαι τρελός ("seipazzo")
9. Cos'hai combinato? - Τι έκανες; ("combinato")
10. Nonme ne frega niente - Δεν με νοιάζει καθόλου ("freganiente")
11. OK! - OK! ("ok")
12. Cosa ti farei! - Θα δεις τι θα πάθεις ("cosatifarei")
13. Basta! - Φτάνει πια ("basta")
14. Le vuoi prendere? - Θες να σε χτυπήσω; ("prendere")
15. Non ce n'e piu - Έχει τελειώσει ("noncenepiu")
16. Ho fame - Πεινάω ("fame")
17. Tanto tempo fa - Πολύ καιρό πριν ("tantotempo")
18. Buonissimo - Πολύ νόστιμο ("buonissimo")
19. Si sono messi d'accordo! - Έχουν κάνει μια (μυστική) συμφωνία ("messidaccordo")
20. Sono stufo - Έχω βαρεθεί ("sonostufo")

Παράλληλα με την παρουσίαση του λεξιλογίου, παραπέμπουμε και στην οπτικοποίηση των εν λόγω χειρονομιών που γίνεται στο σχήμα 1.4. Για κάθε μία από τις 20 παραπάνω χειρονομίες, παρουσιάζουμε ένα αντιπροσωπευτικό καρέ από την εκτέλεσή της, εστιάζοντας στη θέση των χεριών, αλλά και στη χειρομορφή που κατά κύριο λόγο τη χαρακτηρίζει.

Να σημειώσουμε εδώ ότι οι εκτελέσεις των εν λόγω χειρονομιών δεν εντάσσονται στα πλαίσια ενός αυστηρού λεξιλογίου με συγκεκριμένους κανόνες πραγματοποίησης (όπως για παράδειγμα συναντάται στη νοηματική γλώσσα), αλλά αντίθετα πρόκειται για χειρονομίες οι οποίες συνοδεύουν αυθόρμητα την ομιλία στην Ιταλική γλώσσα. Συνεπώς, η ακριβής εκτέλεση μπορεί να διαφέρει μεταξύ διαφορετικών χρηστών, ή ακόμα και μεταξύ διαφορετικών εκτελέσεων του ίδιου χρήστη. Παραδείγματα τέτοιων διαφορετικών εκτελέσεων έχουν ήδη παρουσιαστεί στο σχήμα 1.1.

(α′) “vattene” (β′) “vieniqui” (γ′) “perfetto” (δ′) “furbo” (ε′) “cheduepalle”

(ϝ′) “chevuoi” (ζ) “daccordo” (η′) “seipazzo” (θ′) “combinato” (ι′) “freganiente”

(ια′) “ok” (ιβ′) “cosatifarei” (ιγ′) “basta” (ιδ′) “prendere” (ιε′) “noncenepiu”

(ιϝ′) “fame” (ιζ) “tantotempo” (ιη′) “buonissimo” (ιθ′) “messidaccordo” (κ′) “sonostufo”

Σχήμα 1.4: Λεξιλόγιο χειρονομιών της πολυτροπικής βάσης ChaLearn. Παρουσιάζονται χαρακτηριστικά στιγμιότυπα για κάθε χειρονομία, έτσι όπως εκτελούνται από το συγκεκριμένο χρήστη.

Δομή των Δεδομένων

Τα δεδομένα έχουν καταγραφεί και παρέχονται σε συνεδρίες των 60 με 90 δευτερολέπτων. Στο συγκεκριμένο χρονικό διάστημα ο χρήστης υπό εξέταση εκτελεί ένα υποσύνολο των χειρονομιών του λεξιλογίου σημειώνοντας μικρές παύσεις μεταξύ των διαφορετικών εκτελέσεων. Κατά κανόνα δεν υπάρχουν επαναλήψεις της ίδιας χειρονομίας σε κάθε συνεδρία. Από την άλλη όμως είναι πιθανό κάποιες συνεδρίες να έχουν εμπλουτιστεί με χειρονομίες εκτός του λεξιλογίου, αυξάνοντας έτσι τη δυσκολία του έργου της αναγνώρισης. Όπως αναφέρθηκε προηγουμένως, οι συνεδρίες περιλαμβάνουν συνολικά 39 χρήστες, κάθε ένας από τους οποίους εμφανίζεται σε ένα πλήθος από 5 έως και 50 συνεδρίες περίπου. Συγκεντρωτικά, έχουν καταγραφεί πάνω από 13000 εκτελέσεις των χειρονομιών εντός του λεξιλογίου.

Τέλος για τις ανάγκες και του Διαγωνισμού, τα δεδομένα έχουν ταξινομηθεί σε τρία μεγάλα σύνολα:

- Το Σύνολο Εκπαίδευσης ή Training Set, το οποίο περιλαμβάνει 387 συνεδρίες και 7740 εκτελέσεις συνολικά. Στόχο έχει να χρησιμοποιηθεί καθαρά για την εκπαίδευση των συστημάτων αναγνώρισης.

- Το Σύνολο Επικύρωσης ή Validation Set, το οποίο περιλαμβάνει 287 συνεδρίες και 3345 εκτελέσεις χειρονομιών που περιλαμβάνονται στο λεξιλόγιο. Στόχο έχει να χρησιμοποιηθεί για τη ρύθμιση των παραμέτρων της εκπαίδευσης, αφού έχει αντίστοιχη μορφή με το τελικό σύνολο δεδομένων (Test Set). Παράλληλα όμως, μπορεί να χρησιμοποιηθεί για την εκπαίδευση του συστήματος αναγνώρισης

- Το Σύνολο Αξιολόγησης ή Test Set, το οποίο περιλαμβάνει 275 συνεδρίες και 2200 εκτελέσεις χειρονομιών εντός του λεξιλογίου. Πρόκειται για το τελικό σύνολο πάνω στο οποίο και θα αξιολογηθούν τα συστήματα αναγνώρισης.

Συνοψίζοντας, τα ενδιαφέροντα χαρακτηριστικά της συγκεκριμένης βάσης και τα οποία δικαιολογούν την επιλογή της για τα πειράματά μας, είναι:

- Πολυτροπική αναπαράσταση της πληροφορίας

- Πλούσιο λεξιλόγιο (20 χειρονομίες)

- Ποικιλία διαφορετικών χρηστών (39 χρήστες)

- Εκτεταμένος αριθμός συνολικών εκτελέσεων (πάνω από 13000)

1.5.3 Η πολυτροπική και πολυ-αισθητηριακή βάση χειρονομιών MOBOT

Παράλληλα με τη βάση δεδομένων ChaLearn που αποτέλεσε το κύριο αντικείμενο μελέτης, είχαμε στη διάθεσή μας μία ακόμα πολυτροπική, αλλά και πολυ-αισθητηριακή βάση χειρονομιών. Η βάση αυτή εξυπηρετεί τους σκοπούς του ερευνητικού προγράμματος MOBOT, για την κατασκευή μίας ρομποτικής πλατφόρμας υποστήριξης ηλικιωμένων ατόμων, η οποία θα ενσωματώνει και χαρακτηριστικά πολυτροπικής επικοινωνίας [63]. Για το λόγο αυτό, το σύνολο των χρηστών της αποτελείται από άτομα μεγάλης ηλικίας με κινητικά, και πολλές φορές, και διανοητικά προβλήματα.

Στη συγκεκριμένη υποενότητα κάνουμε μια σύντομη παρουσίαση της MOBOT βάσης, στα πλαίσια που αφορούν την πολυτροπική αναγνώριση χειρονομιών. Αν και ο πειραματισμός στη συγκεκριμένη βάση είναι μάλλον πρώιμος, και παρουσιάζεται μικρό κομμάτι ερευνητικών αποτελεσμάτων, έχει ενδιαφέρον να παρατηρήσουμε ότι οι στόχοι του προγράμματος MOBOT αποτελούν μια ρεαλιστική απόδειξη της σημασίας της έρευνας στο πεδίο της αναγνώρισης χειρονομιών, αλλά και των πρακτικών εφαρμογών που μπορούν να υλοποιηθούν.

Παρεχόμενες Τροπικότητες/Αισθητηριακά Σχήματα

Για την εν λόγω βάση δεδομένων υπήρχε η δυνατότητα, αλλά και η ανάγκη, να καταγραφεί ένα πλήθος διαφορετικών τροπικοτήτων, από μια σειρά από διαφορετικούς αισθητήρες. Στη συνέχεια απαριθμούμε το σύνολο των διαφορετικών αισθητήρων, μαζί με κάποια σύντομα σχόλια για τα δεδομένα που λάβαμε από καθένα από αυτούς:

- Άνωθεν Kinect: Αισθητήρας Kinect τοποθετημένος πάνω στη ρομποτική πλατφόρμα, με σκοπό την καταγραφή του άνω μέρος του σώματος του χρήστη. Παρείχε δεδομένα RGB και βάθους.

- Κάτωθεν Kinect: Αισθητήρας Kinect τοποθετημένος πάνω στη ρομποτική πλατφόρμα, με σκοπό την καταγραφή του κάτω μέρος του σώματος του χρήστη. Παρείχε δεδομένα RGB και βάθους.

- Κάμερα GoPro: Κάμερα ευρείας γωνίας λήψης τοποθετημένη πάνω στη ρομποτική πλατφόρμα, με σκοπό την καταγραφή του άνω μέρους του σώματος του χρήστη. Παρείχε δεδομένα RGB υψηλής ευκρίνειας.

- Συστοιχία μικροφώνων MEMS: Σειρά μικροφώνων τεχνολογίας MEMS τοποθετημένα σε γραμμική συστοιχία πάνω στη ρομποτική πλατφόρμα. Παρέχουν πολυκαναλική (8 κανάλια) καταγραφή δεδομένων ήχου.

- Δύο κάμερες υψηλής ευκρίνειας: Σταθερές κάμερες υψηλής ευκρίνειας, για την καταγραφή ολόκληρου του πεδίου δράσης. Παρέχουν δεδομένα RGB υψηλής ευκρίνειας.

- Οπτικό σύστημα καταγραφής σκελετού: Σύστημα τύπου Qualisys Motion Capture System που επιτρέπει την καταγραφή του σκελετού του χρήστη μεσω ειδικών markers που είναι τοποθετημένοι σε συγκεκριμένα σημεία στο σώμα του χρήστη.

Παράλληλα, στην καταγραφή δεδομένων ήταν διαθέσιμοι κάποιοι επιπλέον αισθητήρες, όπως καταγραφείς laser, αισθητήρες δύναμης-ροπής και άλλοι. Ωστόσο, εδώ παρουσιάσαμε μόνο όσους θα μπορούσαν να χρησιμοποιηθούν για τις ανάγκες της πολυτροπικής αναγνώρισης χειρονομιών, δηλαδή αυτούς που παρέχουν οπτική και ηχητική πληροφορία.

Λεξιλόγιο

Για τις ανάγκες της βάσης δημιουργήθηκε ένα εκτενές λεξιλόγιο χειρονομιών, με σκοπό την υποστήριξη της επικοινωνίας των ηλικιωμένων ασθενών με τη ρομποτική πλατφόρμα. Κάθε εντολή προς την πλατφόρμα αρχίζει με το όνομα 'MOBOT' (για να είναι διακριτή η επιθυμία επικοινωνίας με την πλατφόρμα) και συνεχίζει με την εκφορά της ζητούμενης εντολής. Το πλήρες λεξιλόγιο παρουσιάζεται στη συνέχεια, και συνοδεύεται από τις εντολές, καθώς και την απεικόνιση των αντίστοιχων χειρονομιών (σχήμα 1.5).

1. "Help" - Βοήθεια
2. "I want to stand up" - Θέλω να σηκωθώ
3. "I want to perform a task" - Θέλω να εκτελέσω μία εργασία
4. "I want to sit down" - Θέλω να κάτσω
5. "Come here" - Έλα εδώ
6. "Come closer" - Έλα πιο κοντά
7. "Go straight" - Πήγαινε ευθεία
8. "Park" - Πάρκαρε
9. "Stop" - Σταμάτα
10. "Go away" - Φύγε

11. “Let’s go” - Ξεκινάμε
12. “Turn left” - Στρίψε αριστερά
13. “Turn Right” - Στρίψε Δεξιά
14. “Avoid an obstacle” - Απέφυγε το εμπόδιο
15. “I want to go through the door” - Θέλω να περάσω από την πόρτα
16. “Yes” - Ναι
17. “No” - Όχι
18. “Where am I?” - Που βρίσκομαι;
19. “What time is it?” - Τι ώρα είναι;

Δομή των Δεδομένων

Επικεντρώνοντας στο σενάριο καταγραφής που έχει να κάνει αποκλειστικά με την πραγματοποίηση χειρονομιών, έχουν καταγραφεί δεδομένα από 13 ηλικιωμένους ασθενείς συνολικά. Κάθε ένας από αυτούς εκτελεί όλες τις χειρονομίες του παραπάνω λεξιλογίου (σε συνδυασμό με τις αντίστοιχες φωνητικές εντολές), πραγματοποιώντας τρεις έως πέντε εκτελέσεις σε κάθε περίπτωση.

Πειραματισμοί

Για τις ανάγκες των πειραματισμών που θα παρουσιαστούν στη συνέχεια, χρησιμοποιήθηκε ένα υποσύνολο των διαθέσιμων ασθενών (8 από τους 13), αλλά και των διαθέσιμων χειρονομιών (8 από τις 19 διαθέσιμες, και συγκεκριμένα οι 1 έως 8 της προηγούμενης λίστας). Επίσης, αξιοποιήθηκαν δεδομένα μόνο από το άνωθεν Kinect, που εστιάζει στον κορμό του χρήστη, ενώ χρησιμοποιήθηκε μια επισημείωση των σημείων του σκελετού, που προέκυψε από παρατήρηση των καρέ του βίντεο.

Δυσκολίες/Προκλήσεις

Εν μέρει λόγω της ιδιαιτερότητας των χρηστών που πραγματοποιούν τις χειρονομίες στη συγκεκριμένη περίπτωση, η MOBOT βάση παρουσιάζει κάποιες επιπλέον δυσκολίες που μπορεί να συναντώνται και σε άλλες βάσεις, όμως εδώ είναι ακόμα πιο έντονες. Πιο συγκεκριμένα, έχουμε:

Σχήμα 1.5: Λεξιλόγιο χειρονομιών της βάσης δεδομένων MOBOT. Παρουσιάζονται χαρακτηριστικά στιγμιότυπα για κάθε χειρονομία, έτσι όπως εκτελούνται από το συγκεκριμένο χρήστη. (Έχει σκόπιμα εφαρμοστεί μία θόλωση στα πρόσωπα, για λόγους προστασίας προσωπικών δεδομένων).

Χρήση τεχνητού λεξιλογίου. Συνήθως οι χειρονομίες που πραγματοποιούμε αφορούν ένα μέρος του συνηθισμένου μας λεξιλογίου (π.χ. "έλα εδώ" ή "φύγε"), οπότε προκύπτουν πολύ φυσικά στην εκτέλεσή τους. Αντίθετα, στη συγκεκριμένη περίπτωση έχουμε δημιουργήσει ένα τεχνητό λεξιλόγιο για συγκεκριμένες ανάγκες επικοινωνίας με τη ρομποτική πλατφόρμα, οπότε σε ορισμένες περιπτώσεις είναι πιθανό οι χειρονομίες να μην προκύπτουν το ίδιο φυσικά, και να είναι πιο δύσκολο για τους χρήστες να τις πραγματοποιήσουν στα πλαίσια ενός αληθινού σεναρίου.

Διαφορετικές εκφορές μεταξύ των χρηστών. Είναι πολύ συνηθισμένο να υπάρχουν εναλλακτικές "εκφορές" της ίδιας χειρονομίας από διαφορετικούς ανθρώπους, όπως συζητήθηκε και νωρίτερα (ενότητα 1.1), ωστόσο στη συγκεκριμένη περίπτωση αυτό το φαινόμενο είναι ακόμα πιο έντονο από ότι συνήθως. Λόγω των κινητικών και άλλων παθολογικών προβλημάτων των ηλικιωμένων ασθενών, είναι σύνηθες να περιορίζονται σωματικά, και να εκτελούν τις χειρονομίες με διαφορετικό τρόπο ο καθένας.

Διαφορετικές εκφορές για τον ίδιο χρήστη. Λόγω νοητικών υστερήσεων σε κάποιους ηλικιωμένους ασθενείς της συγκεκριμένης καταγραφής, έχει επίσης παρατηρηθεί σε μερικές περιπτώσεις ότι πραγματοποιούνται διαφορετικές "εκφορές" της ίδιας χειρονομίας από τον ίδιο χρήστη. Κατά κάποιο τρόπο δηλαδή δεν υπάρχει η αναμενόμενη συνέπεια που παρατηρείται στις κινήσεις σωματικώς και διανοητικώς υγιών ατόμων.

1.6 Συνεισφορές της Διπλωματικής Εργασίας

Ολοκληρώνοντας το κεφάλαιο της εισαγωγής, επιχειρούμε να κάνουμε μια σύνοψη των συνεισφορών της παρούσας διπλωματικής εργασίας. Ασφαλώς και η εικόνα αυτή θα είναι πιο ξεκάθαρη μετά την παρουσίαση και την κάλυψη όλων των θεμάτων που διαπραγματεύεται η εργασία, ωστόσο αυτή η ενότητα μπορεί να λειτουργήσει σαν πρόλογος των όσων θα ακολουθήσουν. Ως κυριότερες συνεισφορές της διπλωματικής αυτής, θεωρούμε τις παρακάτω:

- Αναλυτική εξέταση του προβλήματος της πολυτροπικής αναγνώρισης χειρονομιών, τόσο μέσω επεξεργασία οπτικών δεδομένων, όσο και με τη σύμμειξη των οπτικών καναλιών πληροφορίας με την τροπικότητα του ήχου.

- Πειραματισμός σε τρεις διαφορετικές βάσεις, μία βάση στατικών χειρομορφών, μία πολυτροπική βάση χειρονομιών, και μία πολυτροπική και πολυ-αισθητηριακή βάση χειρονομιών, εστιασμένη σε ηλικιωμένους χρήστες.

- Χρήση ισχυρών εργαλείων από την αναγνώριση προτύπων για τη μοντελοποίηση και την αναγνώριση των χειρονομιών, όπως τα Κρυφά Μαρκοβιανά Μοντέλα (HMMs), αλλά και άλλων ταξινομητών όπως Support Vector Machines (SVM) και k-Nearest Neighbor (kNN).

- Αξιοποίηση του καναλιού πληροφορίας της χειρομορφής του χρήστη που χειρονομεί, εξαγωγή πλήθους διαφορετικών οπτικών περιγραφητών, και εκτενής πειραματισμός.

- Αξιοποίηση του καναλιού πληροφορίας της θέσης-κίνησης, με χρήση χαρακτηριστικών που αφορούν τη θέση (σχετική θέση, απόσταση) και την κίνηση (ταχύτητα, επιτάχυνση) των χεριών και των αγκώνων. Αντίστοιχα, ακολούθησε εκτενής πειραματισμός και για τη χρήση των συγκεκριμένων χαρακτηριστικών.

- Μελέτη μεθόδων σύμμειξης των διαφορετικών καναλιών πληροφορίας και τροπικοτήτων, και πρόταση συνδυασμού τους.

- Επίτευξη ποσοστών επιτυχίας στην πολυτροπική βάση χειρονομιών ChaLearn [27], που ξεπερνούν τις επιδόσεις που σημειώθηκαν στον πρόσφατο διαγωνισμό πολυτροπικής αναγνώρισης χειρονομιών [28].

Κεφάλαιο 2

Υπόβαθρο (Background)

Στο κεφάλαιο αυτό, δίνουμε το απαραίτητο υπόβαθρο πριν περάσουμε στην πιο στοχευμένη προσπάθεια που έγινε στο πρόβλημα της αναγνώρισης χειρονομιών. Πιο συγκεκριμένα, φιλοδοξούμε να εστιάσουμε σε κάποια βασικά σημεία που χαρακτηρίζουν τις μεθόδους και τις τεχνικές που αναπτύξαμε· τόσο από την πλευρά της παρουσίασης του θεωρητικού υποβάθρου, όσο και από την αποσαφήνιση λεπτομερειών της υλοποίησής μας. Εφόσον στα κεφάλαια που ακολουθούν θα περιγράψουμε αναλυτικά τις ερευνητικές μας προσπάθειες στην εξαγωγή χαρακτηριστικών, και στην πολυτροπική σύμμειξη, εκμεταλλευόμαστε την ευκαιρία να παρουσιάσουμε κάποιες βασικές έννοιες από την αναγνώριση προτύπων, αλλά και κάποιες μεθοδολογίες που χρησιμοποιήθηκαν εκτενώς στην παρούσα διπλωματική, για τις διαδικασίες της ταξινόμησης και της αναγνώρισης.

Καλό είναι να σημειώσουμε ότι δεν πρόκειται για πλήρη παρουσίαση του συστήματος που χρησιμοποιούμε, αλλά των βασικών εννοιών, μεθόδων και τακτικών που είναι κοινές σε όλη την έκταση της διπλωματικής, όπως παράδειγμα η μεθοδολογία εκπαίδευσης και αναγνώρισης που εφαρμόζουμε.

2.1 Κρυφά Μαρκοβιανά Μοντέλα (HMMs)

Τα Κρυφά Μαρκοβιανά Μοντέλα, ή HMMs, αποτέλεσαν τη βάση του πειραματισμού μας στο πρόβλημα της αναγνώρισης χειρονομιών, όπως θα φανεί και στη συνέχεια. Η μεγάλη δύναμη των HMMs είναι η δυνατότητα περιγραφής ενός δυναμικά εξελισσόμενου φαινομένου, και η στατιστική του μοντελοποίηση. Σε αντίθεση με άλλα εργαλεία της αναγνώρισης προτύπων και της μηχανικής μάθησης (όπως για παράδειγμα τα SVM ή τα kNN που θα μας απασχολήσουν στη συνέχεια), τα HMMs μοντελοποιούν την εξέλιξη ενός φαινομένου σε σχέση με το χρόνο, κάνοντάς τα μία "φυσική" επιλογή για ένα πρόβλημα όπως αυτό της αναγνώρισης χειρονομιών.

Στις υποενότητες που ακολουθούν, δεν επιχειρούμε μια πλήρη παρουσίαση

των HMMs (για αυτό παραπέμπουμε σε κάποιες κλασικές πηγές [67, 68]). Αντ' αυτού, προχωράμε αρχικά σε μία αναφορά της βασικής ορολογίας των HMMs, για λόγους πληρότητας και ευκολίας παρακολούθησης του υπολοίπου της διπλωματικής εργασίας, και στη συνέχεια παρουσιάζουμε πιο εξειδικευμένα τον τρόπο αξιοποίησης των HMMs με στόχο την αναγνώριση χειρονομιών.

2.1.1 Θεωρητικό Υπόβαθρο

Ένα Κρυφό Μαρκοβιανό Μοντέλο αποτελεί μια μηχανή πεπερασμένων καταστάσεων, οι οποίες δεν είναι άμεσα παρατηρήσιμες παρά μόνο μέσω της ακολουθίας των παραγόμενων συμβόλων του μοντέλου. Για την περιγραφή ενός HMM απαιτείται ο ορισμός ενός συνόλου παραμέτρων:

- Ένα HMM αποτελείται από ένα σύνολο N διαφορετικών καταστάσεων, και σε κάθε χρονική στιγμή $t = 1, 2, ...T$ το μοντέλο βρίσκεται σε μία από αυτές $q_t = \{q_1, q_2, ..., q_N\}$.

- Για την μετάβαση από μία κατάσταση i σε μία κατάσταση j ορίζεται η πιθανότητα μετάβασης: $\mathbf{A} = \{a_{ij}\} : a_{ij} = Pr(q_{t+1} = j | q_t = i), \quad 1 \leq i, j \leq N$. Οι πιθανότητες a_{ij} θα πρέπει να ικανοποιούν τις σχέσεις:

$$a_{ij} \geq 0 \tag{2.1}$$

$$\sum_{j=1}^{N} a_{ij} = 1, \forall i \tag{2.2}$$

- Σε κάθε κατάσταση εξάγεται μια παρατήρηση $\mathbf{o}_t$ από το σύνολο συμβόλων V των δυνατών παρατηρήσεων. Σε κάθε κατάσταση j αντιστοιχεί μια διαφορετική συνάρτηση κατανομής για την πιθανότητα εξαγωγής του κάθε συμβόλου: $\mathbf{B} = \{b_j\} : b_j(\mathbf{o}_t) = Pr(\mathbf{o}_t | q_t = j)$. Στην περίπτωση HMM με συνεχή συνάρτηση πυκνότητας πιθανότητας αυτή θα ορίζεται σαν το συνδυασμό M Γκαουσιανών κατανομών:

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{M} c_{jk} \mathcal{N}(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \mathbf{U}_{jk}), 1 \leq j \leq N \tag{2.3}$$

  Οι παράγοντες $c_{jk}, \boldsymbol{\mu}_{jk}, \mathbf{U}_{jk}$ αντιπροσωπεύουν τους συντελεστές των Γκαουσιανών, το διάνυσμα των μέσων όρων και τον πίνακα συμμεταβλητότητας αντίστοιχα για την k κατανομή και την j κατάσταση. Επίσης, οι συντελεστές των Γκαουσιανών θα πρέπει να ικανοποιούν τους

παρακάτω περιορισμούς:

$$\sum_{k=1}^{M} c_{jk} = 1, \quad 1 \leq j \leq N \tag{2.4}$$

$$c_{jk} \leq 0, \quad 1 \leq j \leq N, \quad 1 \leq k \leq M \tag{2.5}$$

- Τέλος, ορίζεται και μια κατανομή πιθανότητας π_i η οποία εκφράζει την πιθανότητα το HMM να ξεκινά από την κατάσταση i:
  $\boldsymbol{\pi} = \{\pi_i\} : \pi_i = Pr(q_1 = i), 1 \leq i \leq N$.

Έτσι, με βάση τα παραπάνω, ένα HMM μπορεί να περιγραφεί πλήρως από το σύνολο παραμέτρων $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$.

Έχοντας ολοκληρώσει τη βασική περιγραφή των HMM, η χρήση τους σε προβλήματα της αναγνώρισης προτύπων γίνεται εφικτή με τη χρήση κάποιων βασικών αλγορίθμων που επιτρέπουν την εκπαίδευση αλλά και την αναγνώριση στα HMMs. Στη συνέχεια κάνουμε μια σύντομη αναφορά σε καθένα από αυτούς.

- Οι αλγόριθμοι forward και backward είναι δύο ιδιαίτερα αποδοτικοί αλγόριθμοι, οι οποίοι δεδομένης μιας ακολουθίας παρατηρήσεων $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T)$ και ενός HMM με παραμέτρους $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ επιτρέπουν τον υπολογισμό της πιθανοφάνειας $Pr(\mathbf{O}|\boldsymbol{\lambda})$.

- Ο αλγόριθμος Viterbi για μια δεδομένη ακολουθία καταστάσεων $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T)$ και ενός μοντέλου $\boldsymbol{\lambda}$, μας επιτρέπει να βρούμε την ακολουθία καταστάσεων $Q = q_1, q_2, ..., q_T$ που μεγιστοποιεί την πιθανότητα $Pr(\mathbf{O}, \mathbf{q}|\boldsymbol{\lambda})$.

- Ο αλγόριθμος Baum-Welch, που πρόκειται για ειδική περίπτωση του EM (Expectation-Maximization), χρησιμοποιείται για την προσεγγιστική εύρεση των βέλτιστων παραμέτρων που μεγιστοποιούν την πιθανοφάνεια του μοντέλου $Pr(\mathbf{O}|\boldsymbol{\lambda})$ για ένα σύνολο εκπαίδευσης. Το πρόβλημα αυτό είναι και το πιο δύσκολο (δεν υπάρχει αναλυτική λύση), και αντιστοιχεί στην εκπαίδευση των HMMs, δεδομένου ενός συνόλου εκπαίδευσης.

2.1.2 Λεπτομέρειες Υλοποίησης

Στη συνέχεια, με γνωστό το βασικό θεωρητικό υπόβαθρο των HMMs, περιγράφουμε κάποιες από τις βασικές λεπτομέρειες της υλοποίησης, κατά την εργασία μας με τα HMMs. Ο σκοπός της συγκεκριμένης υποενότητας είναι να αποσαφηνίσουμε κάποια σημεία που αφορούν τις μεθόδους που χρησιμοποιήσαμε. Συνήθως έχουν να κάνουν με τη μεθοδολογία που ακολουθήσαμε, με παραδοχές που κάναμε, με χαμηλού επιπέδου και γενικότερα μικρές αλλά σημαντικές λεπτομέρειες οι

οποίες δεν αναφέρονται ρητά στα επόμενα κεφάλαια, αλλά είναι απαραίτητες για την κατανόηση του τρόπου εργασίας μας.

Σημειώνουμε εδώ, ότι για την υλοποίηση των HMMs χρησιμοποιήσαμε το γνωστό toolbox htk [103], το οποίο παρέχει υλοποιήσεις των βασικών αλγορίθμων για HMMs και αποτελεί ένα ολοκληρωμένο περιβάλλον χρήσης HMMs.

Εκπαίδευση

Για την εκπαίδευση των μοντέλων μας χρησιμοποιήθηκαν τα εργαλεία HRest και HERest του htk, τα οποία αποτελούν υλοποιήσεις του αλγορίθμου Baum-Welch. Σε κάθε περίπτωση, με δεδομένο το σύνολο εκπαίδευσης, όλα τα παραδείγματα που εντάσσονταν σε αυτό, χρησιμοποιήθηκαν για την εκπαίδευση του συστήματος. Η διαδικασία της εκπαίδευσης περιελάμβανε την εκπαίδευση ενός μοντέλου για κάθε χειρονομία-λέξη του λεξιλογίου μας. Στις περιπτώσεις δε, που αντιμετωπίσαμε προβλήματα αναγνώρισης (περισσότερα για αυτά παρουσιάζουμε στην υποενότητα της *Αξιολόγησης*), προχωρήσαμε στην εκπαίδευση ενός ακόμα μοντέλου, για το οποίο χρησιμοποιήσαμε τα διαθέσιμα παραδείγματα εκπαίδευσης από όλες τις κλάσεις. Σκοπός αυτού του μοντέλου (που στη συνέχεια το αναφέρουμε ως garbage ή background model) ήταν να εντοπίσουμε αποσπάσματα της ακολουθίας, τα οποία είτε δεν περιλαμβάνουν κάποια χειρονομία, είτε περιλαμβάνουν χειρονομίες εκτός του λεξιλογίου εκπαίδευσης.

Ο αριθμός καταστάσεων αλλά και το πλήθος των Γκαουσιανών ανά κατάσταση χαρακτηρίζει τα μοντέλα που εκπαιδεύσαμε και γι᾽ αυτό αναφέρεται ρητά σε κάθε περίπτωση, στα πειράματα που εκτελέσαμε στα επόμενα κεφάλαια. Ασφαλώς, λόγω της επιρροής που έχουν αυτές οι δύο παράμετροι στη σωστή μοντελοποίηση των χειρονομιών, δεν τέθηκαν σταθερές σε όλες τις περιπτώσεις, αλλά αποτέλεσαν αντικείμενο συχνού πειραματισμού. Ωστόσο, όπου στη συνέχεια πραγματοποιούνται συγκρίσεις μεταξύ μεθόδων ή τεχνικών, φροντίσαμε να έχουμε κοινό αριθμό καταστάσεων, ώστε να προκύψουν και αντικειμενικά συμπεράσματα.

Τέλος, όσον αφορά την τοπολογία των HMMs, επιλέξαμε να χρησιμοποιήσουμε αποκλειστικά left-to-right μοντέλα (και πιο συγκεκριμένα μοντέλα που επιτρέπουν παραμονή στην ίδια κατάσταση, ή μετάβαση μόνο στην επόμενη κατάσταση, αντί για οποιαδήποτε επόμενη). Αυτή η επιλογή έγινε, γιατί θεωρούμε ότι τα left-to-right μοντέλα αποτελούν τη “φυσική” επιλογή για ένα φαινόμενο που εξελίσσεται στο χρόνο, χωρίς να έχουμε επιστροφή σε προηγούμενες καταστάσεις. Τα left-to-right μοντέλα άλλωστε είναι και η πλέον συνηθισμένη επιλογή για μοντελοποίηση χρονικών σημάτων, όπως για παράδειγμα η φωνή.

Αξιολόγηση

Πριν προχωρήσουμε στην ανάλυση της προσέγγισής/υλοποίησής μας για την αξιολόγηση του συστήματος, είναι σημαντικό να κάνουμε το διαχωρισμό μεταξύ προβλημάτων ταξινόμησης και προβλημάτων αξιολόγησης, δύο έννοιες που θα συναντάμε συνεχώς στα επόμενα κεφάλαια της παρούσας διπλωματικής εργασίας.

- Σε προβλήματα ταξινόμησης εξετάζουμε συγκεκριμένα αποσπάσματα, τα οποία περιλαμβάνουν μία λέξη-χειρονομία του λεξιλογίου εκπαίδευσης, και σκοπός είναι να αναθέσουμε επιτυχημένα κάθε τέτοιο απόσπασμα στην αντίστοιχη κλάση λέξης-χειρονομίας. Το σύνολο των αποσπασμάτων προς ταξινόμηση αποτελεί το σύνολο αξιολόγησης, ενώ η επιτυχία της ταξινόμησης αξιολογείται με βάση το πλήθος των αποσπασμάτων που ταξινομήθηκαν σωστά (H), ως προς το συνολικό αριθμό αποσπασμάτων που εξετάστηκαν (N). Δηλαδή: $Accuracy = \frac{H}{N}$.

- Σε προβλήματα αναγνώρισης από την άλλη πλευρά, εξετάζουμε αποσπάσματα που περιλαμβάνουν αυθαίρετο αριθμό από χειρονομίες. Συνεπώς, σκοπός μας είναι να εντοπίσουμε ποιες χειρονομίες πραγματοποιήθηκαν, αλλά και με ποια σειρά. Όπως αναφέραμε και προηγουμένως, ένα ακόμα μοντέλο χρησιμοποιείται (background), με την φιλοδοξία να αναγνωρίσει τμήματα τα οποία είτε δεν περιλαμβάνουν χειρονομίες, είτε περιλαμβάνουν χειρονομίες εκτός του λεξιλογίου εκπαίδευσης. Ασφαλώς πρόκειται για ένα πρόβλημα δυσκολότερο από αυτό της ταξινόμησης, μιας και απαιτείται τόσο ο εντοπισμός μιας χειρονομίας στη συνεχή ακολουθία, όσο και η ταξινόμησή της στη σωστή κλάση. Για την αξιολόγηση της επιτυχίας της αναγνώρισης, υπολογίζονται το πλήθος των σωστών αναγνωρίσεων (H), των διαγραφών (D), των αντικαταστάσεων (S) και των εισαγωγών (I) της ακολουθίας που αναγνωρίστηκε σε σχέση με την πραγματική. Αν η πραγματική ακολουθία περιλαμβάνει συνολικά N χειρονομίες προς αναγνώριση, τότε η ακρίβεια της αναγνώρισης υπολογίζεται ως: $Accuracy = \frac{H-D-S-I}{N}$.

Σχετικά με την υλοποίησή μας, και στις δύο περιπτώσεις, της ταξινόμησης και της αναγνώρισης, χρησιμοποιήθηκε το εργαλείο HVite του htk, το οποίο υλοποιεί τον αλγόριθμο Viterbi. Η μόνη διαφορά έγκειται στο γεγονός ότι για την περίπτωση της αναγνώρισης χρησιμοποιήθηκε μία γραμματική, που επέτρεπε την αναγνώριση οποιουδήποτε αριθμού χειρονομιών στην ακολουθία. Κατ' ουσίαν, πρόκειται για μία επέκταση του αλγορίθμου Viterbi, η οποία είναι πολύ συνηθισμένη στην αναγνώριση συνεχούς λόγου.

2.2 Οι ταξινομητές SVM και kNN

Σε συνδυασμό με τα HMMs κάναμε χρήση και "στατικών" (σε σύγκριση με τα HMMs) ταξινομητών, όπως τα SVM και kNN, σε πολύ μικρότερη όμως κλίμακα. Τα αποτελέσματα, όπως θα φανεί και στα επόμενα κεφάλαια, ήταν ικανοποιητικά, ωστόσο είμαστε περιορισμένοι σε προβλήματα ταξινόμησης, μιας και δεν υπάρχει κάποια προφανής μέθοδος που να επιτρέπει την επέκταση τέτοιων ταξινομητών σε συνεχή αναγνώριση, όπως συμβαίνει με τα HMMs. Στη συνέχεια κάνουμε μια σύντομη παρουσίαση των ταξινομητών αυτών, αλλά και της μεθοδολογίας χρήσης τους.

2.2.1 Support Vector Machines (SVM)

Τα SVM αποτελούν ταξινομητές μηχανικής μάθησης, που εστιάζουν στον καλύτερο δυνατό διαχωρισμό μεταξύ κλάσεων. Στην ουσία, πρόκειται για γραμμικό ταξινομητή, που φιλοδοξεί να εντοπίσει τη βέλτιστη διαχωριστική γραμμή (για δισδιάστατο χώρο) στο χώρο χαρακτηριστικών, μεταξύ δύο κλάσεων. Για τη βελτιστοποίηση της τοποθέτησης της διαχωριστικής γραμμής χρησιμοποιείται η απλή ιδέα της μεγιστοποίησης της απόστασης των προτύπων κάθε κλάσης από την εν λόγω γραμμή. Φυσικά για χώρους μεγαλύτερων διαστάσεων δεν προκύπτουν απλώς γραμμές, αλλά υπερεπίπεδα, ή σύνολα από υπερεπίπεδα. Για μη γραμμικά διαχωρίσιμες κλάσεις, υπάρχουν μη γραμμικές επεκτάσεις των SVM, που προκύπτουν με έναν απλό μετασχηματισμό ("τέχνασμα του πυρήνα").

2.2.2 k-Nearest Neighbor (kNN)

Ο k-NN είναι ακόμα ένας ισχυρός ταξινομητής της αναγνώρισης προτύπων, τον οποίο χρησιμοποιήσαμε για τους πειραματισμούς μας. Η φιλοσοφία του k-NN βασίζεται στην ταξινόμηση με βάση τα πλέον γειτονικά παραδείγματα εκμάθησης. Συνεπώς, ένα αντικείμενο θα αντιστοιχηθεί σε κάποια κλάση, βασιζόμενο στην πλειοψηφική ψήφο των k κοντινότερων γειτόνων του στο χώρο χαρακτηριστικών. Αν η τιμή του k γίνει ίση με 1, τότε η ανάθεση γίνεται απλά με βάση τον κοντινότερο γείτονα.

2.2.3 Η τεχνική Bag-of-Features

Η παρουσίαση μιας τεχνικής που αφορά την εξαγωγή και την επεξεργασία χαρακτηριστικών, ίσως να μοιάζει παράταιρη σε αυτό το σημείο μεταξύ διαφόρων ταξινομητών μηχανικής μάθησης. Ωστόσο επειδή αυτή η τεχνική είναι που μας επιτρέπει να χρησιμοποιήσουμε τους "στατικούς" ταξινομητές που έχουμε

αναφέρει παραπάνω για ένα δυναμικό πρόβλημα όπως αυτό της αναγνώρισης χειρονομιών, επιλέγουμε να κάνουμε μια σύντομη παρουσίασή της εδώ.

Η μέθοδος Bag-of-Features είναι εμπνευσμένη από την τεχνική Bag-of-Words που είναι δημοφιλής στην επεξεργασία κειμένου. Η ανάγκη από την οποία προκύπτει η εφαρμογή της, οφείλεται στην επιθυμία μας να αναπαραστήσουμε ένα σύνολο διαθέσιμων βίντεο, τα οποίο μπορεί να έχουν διαφορετική διάρκεια, σε ένα χώρο χαρακτηριστικών, με σταθερή διάσταση. Στόχος της συνεπώς είναι να συμπυκνώσει το σύνολο των περιγραφητών κάθε βίντεο σε ένα ιστόγραμμα, κάθε ράβδος του οποίου αποτελεί μία οπτική λέξη. Για την εύρεση του λεξιλογίου, χρησιμοποιείται ο αλγόριθμος k-means επί του συνόλου των χαρακτηριστικών (για όλα τα διαθέσιμα βίντεο), μετά την εφαρμογή του οποίου, θα έχουν προκύψει k οπτικές λέξεις. Σε αυτό το σημείο, η αναπαράσταση Bag-of-Features για κάθε βίντεο, προκύπτει από την ψηφοφορία των χαρακτηριστικών του εν λόγω βίντεο στο συγκεκριμένο ιστόγραμμα οπτικών λέξεων.

Τελικά το ιστόγραμμα που προέκυψε με την Bag-of-Features αναπαράσταση είναι αυτό που χρησιμοποιούμε για την περιγραφή του κάθε βίντεο. Το μέγεθος του ιστογράμματος είναι σταθερό για κάθε βίντεο, και ίσο με το πλήθος των οπτικών λέξεων (k), οπότε μπορεί να χρησιμοποιηθεί ως διάνυσμα χαρακτηριστικών στους SVM και kNN ταξινομητές.

2.2.4 Λεπτομέρειες Υλοποίησης

Εν γένει η αντιμετώπιση της εκπαίδευσης και της αξιολόγησης, στις περιπτώσεις που χρησιμοποιούμε τέτοιου είδους "στατικούς" ταξινομητές, γίνεται με αρκετά ευθύ τρόπο. Το κύριο πρόβλημα της πιθανής διαφοροποίησης στη διάρκεια διαφορετικών ακολουθιών δεδομένων, αντιμετωπίζεται με την τεχνική του Bag-of-Features. Με τα διανύσματα χαρακτηριστικών να έχουν σταθερό μήκος για όλα τα βίντεο-δεδομένα, οι ταξινομητές αναλαμβάνουν τις εργασίες εκμάθησης και αξιολόγησης κατά τα γνωστά. Ανάλογα με το χωρισμό των δεδομένων σε σύνολα εκπαίδευσης και αξιολόγησης, χρησιμοποιούμε τα παραδείγματα του πρώτου συνόλου για την εκπαίδευση του ταξινομητή, ενώ εξετάζεται η δυνατότητα επιτυχημένης ταξινόμησης, με βάση τα δεδομένα του δεύτερου συνόλου. Και πάλι, αξιολογούμε την επίδοση της ταξινόμησης υπολογίζοντας τον λόγο του αριθμού των δεδομένων που αξιολογήθηκαν επιτυχώς, προς το συνολικό αριθμό των δεδομένων που αξιολογήθηκαν.

Οι μόνες ίσως σημαντικές λεπτομέρειες που αξίζει και πρέπει να αναφέρουμε, αφορούν τη χρήση των SVM. Κατ' αρχάς, σημειώνουμε ότι για την υλοποίηση των SVM χρησιμοποιήσαμε τη βιβλιοθήκη LIBSVM [12], που παρέχει ένα ολοκληρωμένο περιβάλλον επιλογών για τη χρήση των SVM. Επίσης, δύο λεπτομέρειες της μεθοδολογίας που αξίζει να αναφέρουμε, είναι οι εξής:

- Η αρχική χρήση των SVM αφορούσε το διαχωρισμό προτύπων δύο, και μόνο, διαφορετικών κλάσεων. Συνεπώς δεν προκύπτει με άμεσο τρόπο πως θα γίνει η επέκτασή τους σε προβλήματα με περισσότερες κλάσεις όπως αυτά που αντιμετωπίζουμε εμείς. Για τέτοιου είδους εφαρμογές, μία συνηθισμένη αντιμετώπιση που έχει προταθεί στη βιβλιογραφία, και χρησιμοποιούμε και εμείς, είναι η τεχνική one-versus-all. Με τη συγκεκριμένη τεχνική, γίνεται εκπαίδευση για κάθε κλάση ξεχωριστά, χρησιμοποιώντας ως αρνητικά παραδείγματα αυτά όλων των υπολοίπων κλάσεων. Στη συνέχεια τα δεδομένα του συνόλου αξιολόγησης εξετάζονται ως προς όλες τις κλάσεις, και η ανάθεση γίνεται στην κλάση για την οποία η συνάρτηση απόφασης του SVM έδωσε τη μεγαλύτερη τιμή.

- Η χρήση μη γραμμικών SVM στη βιβλιογραφία, συνήθως οδηγεί σε βελτιωμένες επιδόσεις σε σχέση με τα γραμμικά SVM. Ειδικά για ιστογραφικά χαρακτηριστικά, όπως αυτά που προκύπτουν από την τεχνική Bag-of-Features, επιλέγεται συνήθως η χ^2-απόσταση, η οποία φαίνεται να αποδίδει καλύτερα σε σχέση με άλλες επιλογές. Συνεπώς και εμείς επεκτείναμε τον πειραματισμό μας και σε μη γραμμικούς SVM ταξινομητές, με χρήση της χ^2-απόστασης.

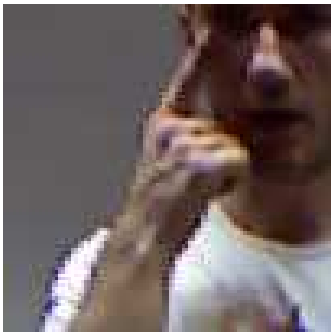
Κεφάλαιο 3

Εκμετάλλευση πληροφορίας Χειρομορφής

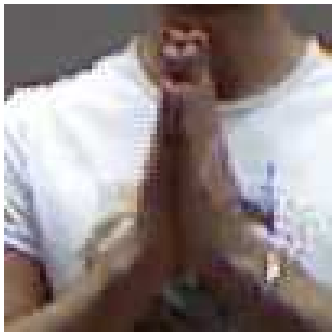
3.1 Γενικά

Το μεγαλύτερο, ίσως, μέρος αυτής της διπλωματικής πραγματεύεται την εκμετάλλευση της οπτικής πληροφορίας για το πρόβλημα της Αναγνώρισης Χειρονομιών. Μια κατεύθυνση θα ήταν αντίστοιχη με αυτές που ακολουθούνται στην Αναγνώριση Ανθρώπινων Δράσεων. Θα μπορούσαμε να εντοπίσουμε χωροχρονικά σημεία ενδιαφέροντος, γειτονιές των οποίων περιλαμβάνουν χρήσιμη πληροφορία. Ή ακόμα θα μπορούσαμε να χρησιμοποιήσουμε τεχνικές πυκνής εξαγωγής χαρακτηριστικών, όπου χρησιμοποιείται όλη η πληροφορία που υπάρχει στη διαθέσιμη εικόνα/βίντεο. Ωστόσο, στην περίπτωση της Αναγνώρισης Χειρονομιών, το πρόβλημα είναι εμφανώς πολύ εστιασμένο για να ακολουθήσουμε τόσο γενικές τεχνικές. Το μεγαλύτερο μέρος της πληροφορίας βρίσκεται στα χέρια, στην εμφάνισή τους, στο σχήμα τους, στις κινήσεις τους. Συνεπώς οριοθετούμε πιο αυστηρά την περιοχή ενδιαφέροντός μας.

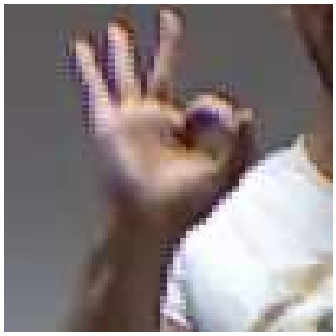
Το κίνητρο αυτό, φαίνεται να επιβεβαιώνεται εν μέρη από την επισκόπηση του σχήματος 3.1. Εδώ έχουμε παρουσιάσει στιγμιότυπα από δύο διαφορετικούς χρήστες να εκτελούν 3 διαφορετικές χειρονομίες. Σκόπιμα έχουμε εστιάσει μόνο στο χέρι (ή στα χέρια) που χειρονομούν. Όπως μπορούμε να δούμε και μόνο η οπτική πληροφορία της χειρομορφής αρκεί για να ταυτοποιήσουμε τις χειρονομίες που εκτελούνται από τους δύο χρήστες. Στη χειρονομία “furbo” όλα τα δάχτυλα του χεριού είναι σε θέση συστολής, εκτός από το δείκτη που έρχεται σε επαφή με το πρόσωπο του χρήστη. Αντίστοιχα στη χειρονομία “combinato”, οι παλάμες των χεριών παραμένουν ενωμένες σε όρθια θέση, κάθετα στον κορμό του χρήστη. Τέλος, για το “ok” ο δείκτης και ο αντίχειρας σχηματίζουν ένα κύκλο, ενώ τα υπόλοιπα τρία δάχτυλα είναι σε έκταση, χωρίς να εφάπτονται.



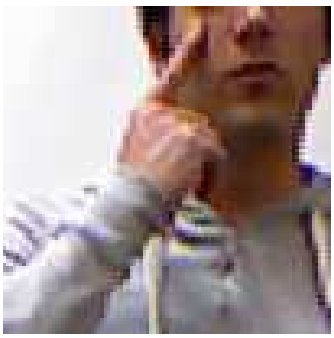
(α') "Furbo" - Χρήστης 1



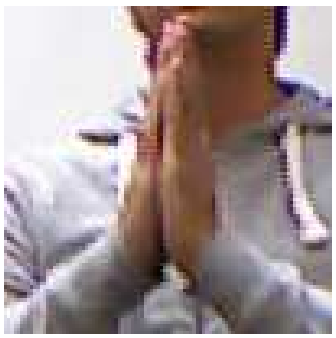
(β') "Combinato" - Χρήστης 1



(γ') "Ok" - Χρήστης 1



(δ') "Furbo" - Χρήστης 2



(ε') "Combinato" - Χρήστης 2

(ϝ') "Ok" - Χρήστης 2

Σχήμα 3.1: Η σημασία της χειρομορφής στην αναγνώριση χειρονομιών. Περιπτώσεις από δύο διαφορετικούς χρήστες, όπου και μόνο η πληροφορία της εμφάνισης της χειρομορφής μπορεί να επιτρέψει την αναγνώριση της εκάστοτε χειρονομίας. Τα στιγμιότυπα προέρχονται από τη βάση χειρονομιών ChaLearn.

Από τα παραπάνω σε καμία περίπτωση δεν πρέπει να υποθέσουμε ότι το πρόβλημα της αναγνώρισης χειρονομιών είναι τόσο απλό όσο αυτό του εντοπισμού και αναγνώρισης κάποιων στατικών χειρομορφών. Ωστόσο, μας δίνει ένα σημαντικό ερέθισμα, ώστε να οριοθετήσουμε την περιοχή ενδιαφέροντος, και να εστιάσουμε εκεί που βρίσκεται το μεγαλύτερο μέρος της πληροφορίας.

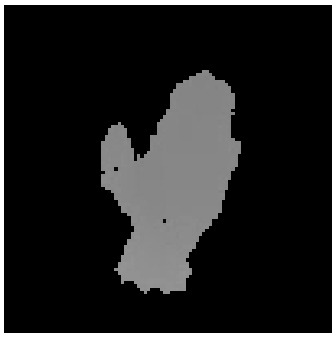
3.2 Μεθοδολογία Εξαγωγής Χαρακτηριστικών

Παρακάτω περιγράφεται ο τρόπος που εφαρμόζουμε για την αξιοποίηση της οπτικής πληροφορίας της χειρομορφής του χρήστη. Σημαντικό ρόλο παίζει όπως θα φανεί και στη συνέχεια, η παρουσία των πλούσιων δεδομένων από το Kinect, με τα οποία καταφέρνουμε να ξεπεράσουμε σχετικά εύκολα κάποια αρκετά απαιτητικά προβλήματα όπως ο εντοπισμός, η παρακολούθηση, αλλά και η κατάτμηση του χεριού.

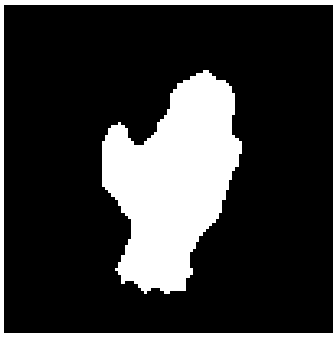
Αρχικά χρησιμοποιούμε το σκελετό που παρέχεται από το Kinect, για τον εντοπισμό της θέσης του χεριού. Με τη χρήση του σκελετού, ο οποίος παρέχεται για κάθε καρέ, πετυχαίνουμε την πλήρη παρακολούθηση του χεριού με την πάροδο του χρόνου. Στη συνέχεια, με γνωστή τη θέση του χεριού επιχειρούμε την κατάτμησή του από το υπόλοιπο σώμα του χρήστη, αλλά και από το παρασκήνιο. Για το σκοπό αυτό, εκμεταλλευόμαστε την πληροφορία του βάθους. Γνωρίζοντας την απόσταση του κεντροειδούς του χεριού από την κάμερα, απομονώνουμε περιοχές με απόσταση εντός ενός ορισμένου κατωφλίου μακρύτερα από αυτή του κεντροειδούς. Αν συνεπώς d_H είναι η απόσταση του κεντροειδούς του χεριού από την κάμερα, η περιοχή που απομονώνουμε απέχει από την κάμερα απόσταση που κυμαίνεται στο εύρος $[d_H - t, d_H + t]$, όπου t το προαναφερθέν κατώφλι. Τέλος, έχοντας εξάγει τη δυαδική μάσκα του χεριού, με χρήση κατωφλιοποίησης στην πληροφορία βάθους, απλά προβάλλουμε αυτή τη μάσκα στο RGB κανάλι, για να προκύψει και εκεί η εικόνα μετά την κατάτμηση της χειρομορφής. Η αναπαράσταση της παραπάνω διαδικασίας παρουσιάζεται και στο σχήμα 3.2, όπου μπορούμε να δούμε την ακρίβεια του αποτελέσματος δεδομένης και της απλότητας της μεθόδου.

Έχοντας εξάγει την εικόνα της χειρομορφής, το επόμενο βήμα είναι η εξαγωγή χαρακτηριστικών σε αυτή την περιοχή. Η εξαγωγή αυτή μπορεί να γίνει είτε στο κανάλι RGB, είτε στο κανάλι βάθους, είτε και στα δύο ταυτόχρονα. Σε κάθε περίπτωση, μπορούμε να χρησιμοποιήσουμε μια πληθώρα περιγραφητών, οι οποίοι είναι γνωστοί και δημοφιλείς στη βιβλιογραφία. Μια αναλυτική λίστα αυτών των περιγραφητών δίνεται στην ενότητα 3.3 μαζί με παρουσίαση του τρόπου υπολογισμού τους. Σε αυτό το σημείο, θα παρουσιάσουμε μόνο μία απεικόνιση από το δημοφιλή περιγραφητή HOG [21] (αναλυτικές πληροφορίες για τον οποίο δίνονται στη συνέχεια), για να δώσουμε την εικόνα της εξαγωγής χαρακτηριστικών

(α') Χειρομορφή (Βάθος)



(β') Κατάτμηση Βάθους



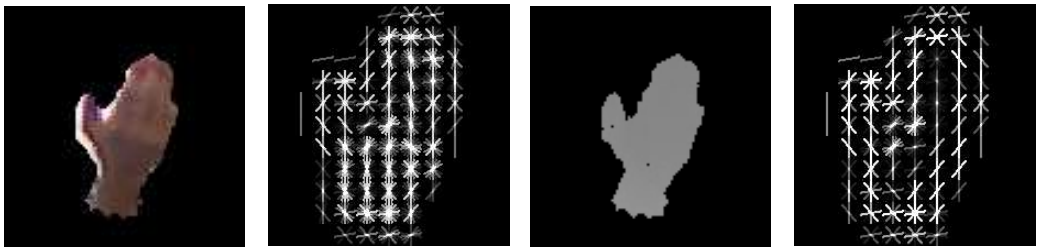
(γ') Μάσκα Χειρομορφής

(δ') Χειρομορφή (RGB)

(ε') Εφαρμογή μάσκας (RGB)

Σχήμα 3.2: Διαδικασία που ακολουθείται για την κατάτμηση της χειρομορφής.

που πραγματοποιούμε. Η οπτικοποίηση του περιγραφητή μετά από εφαρμογή τόσο στο κανάλι της εμφάνισης, όσο και στο κανάλι βάθους, παρουσιάζεται στο σχήμα 3.3



(α') Κανάλι Εμφάνισης (β') HOG Εμφάνισης (γ') Κανάλι Βάθους (δ') HOG Βάθους

Σχήμα 3.3: Εξαγωγή και οπτικοποίηση των χαρακτηριστικών HOG στην περιοχή της χειρομορφής. Χρησιμοποιούνται τόσο η πληροφορία της εμφάνισης, όσο και η πληροφορία βάθους

3.3 Βασικοί οπτικοί περιγραφητές

Σε αυτή την ενότητα γίνεται μια προσπάθεια παρουσίασης των οπτικών περιγραφητών που έχουν χρησιμοποιηθεί στα πλαίσια αυτής της διπλωματικής. Έχοντας εστιάσει αρκετά το πρόβλημά μας, και εργαζόμενοι αποκλειστικά στην περιοχή του χεριού, αντιλαμβανόμαστε ότι είναι απαραίτητο να έχουμε μία όσο το δυνατόν καλύτερη (πλούσια και συμπαγής) αναπαράσταση των χειρομορφών που μελετάμε. Με αυτό το στόχο, παρακάτω περιγράφουμε μια ποικιλία οπτικών περιγραφητών.

Αρχικά, μια πρώτη ταξινόμηση που μπορούμε να κάνουμε είναι με βάση το είδος της πληροφορίας στην οποία επιδρούν αυτοί οι περιγραφητές. Οι κύριες κατηγορίες συνεπώς με βάση το συγκεκριμένο κριτήριο, είναι:

- Οι περιγραφητές *Στατικών Εικόνων*, οι οποίοι επιδρούν σε στατικές εικόνες. Για τις περιπτώσεις που εξετάζουμε, αναγνωρίζουμε δύο κατηγορίες:

  1. Οι περιγραφητές *Εμφάνισης*, που κάνουν εξαγωγή χαρακτηριστικών από την πληροφορία της εμφάνισης, όπως είναι συνήθως το κανάλι RGB ή grayscale.
  2. Οι περιγραφητές *Σχήματος*, που κάνουν εξαγωγή χαρακτηριστικών μόνο από την πληροφορία του σχήματος, δηλαδή ενεργούν πάνω σε δυαδικές εικόνες.

- Οι περιγραφητές *Ακολουθίας Εικόνων* ή περιγραφητές *Κίνησης*, οι οποίοι επιδρούν σε βίντεο. Για το λόγο αυτό, πέρα από την πληροφορία της εμφάνισης, σκοπός τους είναι να ενσωματώσουν και την πληροφορία της κίνησης που υπάρχει στο βίντεο.

Στα πλαίσια της παρούσας διπλωματικής εργασίας χρησιμοποιήθηκε ένα πλήθος διαφορετικών περιγραφητών από κάθε κατηγορία. Συνοπτικά, προκύπτει η επόμενη λίστα:

- Περιγραφητές *Εμφάνισης* (Στατικές Εικόνες)
  1. Histograms of Oriented Gradients (HOG)
  2. Pyramidal HOG (PHOG)
- Περιγραφητές *Σχήματος* (Στατικές Εικόνες)
  1. Fourier Descriptors (FD)
  2. Hu Moments (HU)
  3. Pattern Spectrum (PS)
- Περιγραφητές *Κίνησης* (Ακολουθίες Εικόνων)
  1. Histograms of Optical Flow (HOF)
  2. Συνδυασμός HOG και HOF (HOG/HOF)
  3. HOG3D[1]

Οι παραπάνω περιγραφητές, όσον αφορά τις κατηγορίες *Εμφάνισης* και *Κίνησης*, έχουν χρησιμοποιηθεί στη βιβλιογραφία κυρίως για την εξαγωγή χαρακτηριστικών στην πληροφορία της εμφάνισης (κανάλι RGB ή grayscale). Ωστόσο, μπορούμε το ίδιο εύκολα να τους εφαρμόσουμε στην πληροφορία βάθους, αν την αντιμετωπίσουμε ως μια συνηθισμένη γκρίζα εικόνα. Ένα παράδειγμα αυτής της εξαγωγής χαρακτηριστικών έχει δοθεί ήδη στο σχήμα 3.3δ′, χρησιμοποιώντας τον ιστογραφικό περιγραφητή HOG.

Στη συνέχεια εστιάζουμε στον κάθε ένα από αυτούς τους περιγραφητές ξεχωριστά, περιγράφοντας τον τρόπο υπολογισμού τους. Ιδιαίτερη έμφαση δίνεται στους περιγραφητές HOG, οι οποίοι έχουν χρησιμοποιηθεί και σε μεγαλύτερη έκταση στην παρούσα διπλωματική.

[1] Ο περιγραφητής HOG3D είναι ο μόνος από τους παραπάνω, για τον οποίο δεν παρουσιάζονται πειραματικά αποτελέσματα. Ωστόσο, δίνεται στη λίστα των περιγραφητών, και αναλύεται στη συνέχεια, λόγω του ενδιαφέροντός που παρουσιάζει η ιδέα υπολογισμού του.

HOG

Οι ιστογραφικοί περιγραφητές προσανατολισμένων gradients (HOG), προτάθηκαν για πρώτη φορά από τους Dalal και Triggs το 2005 [21], και έκτοτε παραμένουν από τους πιο δημοφιλείς περιγραφητές σε προβλήματα όπως εντοπισμός και αναγνώριση αντικειμένων, εκτίμηση πόζας, κ.α. αν και η πρώτη τους χρήση ήταν στο πιο εξειδικευμένο πρόβλημα του εντοπισμού ανθρώπων. Μία αναλυτική περιγραφή του αλγορίθμου υπολογισμού τους, έτσι όπως παρουσιάζεται στο [21] δίνεται στη συνέχεια, με σκοπό την παρουσίαση των βασικών σταδίων της διαδικασίας εξαγωγής. Όπως προκύπτει και παρακάτω, πολλές επιλογές κατά την εξαγωγή των HOG χαρακτηριστικών είναι μάλλον αυθαίρετες, ή δεν έχουν κάποια αιτιολόγηση πέρα από το γεγονός ότι επικράτησαν στα αντίστοιχα πειραματικά αποτελέσματα. Εντούτοις, οι HOG περιγραφητές παραμένουν αρκετά δημοφιλείς, και χρησιμοποιούνται ευρέως στη βιβλιογραφία.

Ένα αρχικό στάδιο της αλυσίδας υπολογισμού των HOG, είναι η αναπαράσταση της εικόνας, εξετάζοντας ως διαφορετικές εναλλακτικές τη χρήση τη γκρίζας έκδοσης της εικόνας, τη χρήση του χώρου RGB, ή του χώρου LAB, πιθανώς συνοδευόμενοι από εφαρμογή gamma equalization. Σύμφωνα με τους συγγραφείς οι συγκεκριμένες κανονικοποιήσεις δεν προσφέρουν κάποια ιδιαίτερη βελτίωση, ενώ καταλήγουν στη χρήση του χώρου RGB ή του χώρου LAB, οι οποίοι δίνουν παρόμοια αποτελέσματα.

Στη συνέχεια ακολουθεί ο υπολογισμός των gradients της εικόνας. Και πάλι οι συγγραφείς πειραματίστηκαν με διάφορες εναλλακτικές μάσκες εξαγωγής των gradients, από αρκετά απλές, μέχρι πιο σύνθετες. Σε αυτές συμπεριλαμβάνονται μονοδιάστατες μάσκες (όπως η $[-1, 1]$, η $[-1, 0, 1]$, αλλά και η $[1, -8, 0, 8, -1]$), οι 3×3 μάσκες Sobel, αλλά και 2×2 διαγώνιες μάσκες. Σε αυτή την περίπτωση, προτείνουν την απλή μονοδιάστατη κεντραρισμένη μάσκα $[-1, 0, 1]$, ενώ σε κάθε περίπτωση απορρίπτουν την εφαρμογή κάποιου Γκαουσιανού φιλτραρίσματος ομαλοποίησης, πριν τον εν λόγω υπολογισμό. Από το φιλτράρισμα με τις μονοδιάστατες μάσκες διακριτών παραγώγων $g_x = [-1, 0, 1]$ και $g_y = [-1, 0, 1]^T$, προκύπτουν τα προσανατολισμένα gradients I_x και I_y αντίστοιχα, που επιτρέπουν τον υπολογισμό του μέτρου και της γωνίας του gradient σε κάθε pixel της εικόνας ως εξής:

$$Magnitude(x, y) = \sqrt{{I_x}^2 + {I_y}^2} \tag{3.1}$$

$$Angle(x, y) = \arctan \frac{I_y}{I_x} \tag{3.2}$$

Έχοντας υπολογίσει τα gradients σε κάθε pixel, ακολουθεί το βήμα της χωρικής κβάντισης με βάση την κατεύθυνσή τους. Αρχικά η εικόνα χωρίζεται σε περιοχές-κελιά (cells) δεδομένου μεγέθους (για παράδειγμα τετραγωνικές, 5×5 pixels), και σε κάθε μία από αυτές υπολογίζεται το τοπικό ιστόγραμμα των gradients.

Στο τοπικό ιστόγραμμα, τα bins αντιστοιχούν σε συγκεκριμένο εύρος γωνιών (για παράδειγμα το πρώτο $0^\circ - 20^\circ$, το δεύτερο $20^\circ - 40^\circ$, κοκ), ώστε τα pixels της περιοχής να "ψηφίσουν" στο κατάλληλο bin με βάση την κατεύθυνσή τους. Όσον αφορά την τιμή της "ψήφου", αυτή είναι μια συνάρτηση του μέτρου του gradient, και στην πράξη, αρκεί να ισούται με το ίδιο το μέτρο.

Στο τελευταίο βήμα γίνεται η κανονικοποίηση των τοπικών ιστογραμμάτων, για να αντιμετωπιστούν οι διαφορές στη φωτεινότητα, αλλά και η αντίθεση που μπορεί να υπάρχει σε περιοχές του προσκηνίου και του παρασκηνίου. Για την κανονικοποίηση ορίζονται υπερ-περιοχές κελιών, πιθανώς επικαλυπτόμενες, οι οποίες περιλαμβάνουν περισσότερα του ενός κελιά, και ονομάζονται blocks. Η κανονικοποίηση γίνεται στα τοπικά ιστογράμματα του κάθε block, ενώ με τη συνένωσή τους προκύπτει ο περιγραφητής για το συγκεκριμένο block. Επαναλαμβάνοντας την διαδικασία της κανονικοποίησης για όλα τα blocks, προκύπτει ο τελικός περιγραφητής.

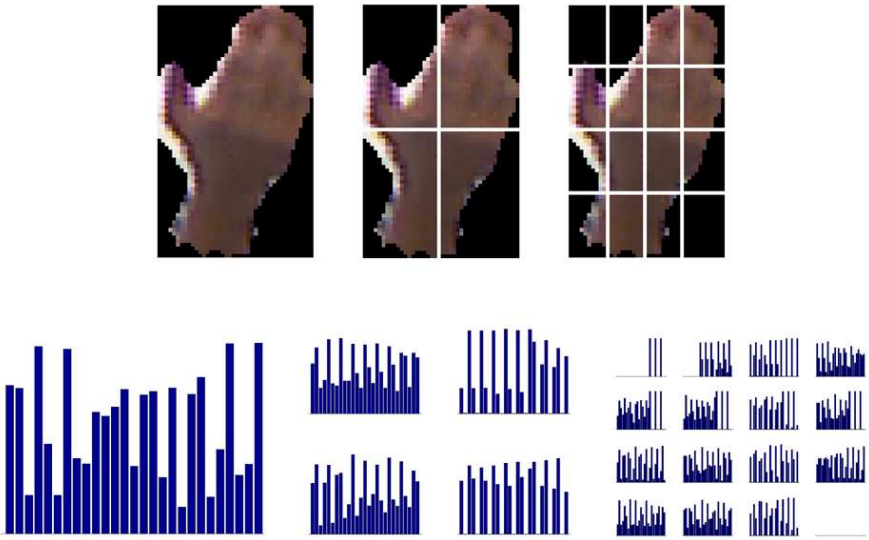
Pyramidal HOG

Η λογική της χρήσης χαρακτηριστικών χρησιμοποιώντας πυραμίδες χωρικής πληροφορίας (Spatial Pyramids), εισήχθη πρώτη φορά από τους Lazebnik et al. [44], και έχει έκτοτε χρησιμοποιηθεί με μεγάλη επιτυχία σε πολλές περιοχές της όρασης υπολογιστών για την εξαγωγή χαρακτηριστικών. Σε συνδυασμό με τα HOG χαρακτηριστικά χρησιμοποιήθηκε για πρώτη φορά από τους Bosch et al. [10].

Η ιδέα πίσω από τους πυραμιδωτούς περιγραφητές χωρικής πληροφορίας είναι η εκμετάλλευση της πληροφορίας σε διαφορετικές κλίμακες της εικόνας. Αντί δηλαδή απλά να υπολογίσουμε τον περιγραφητή στην αρχική εικόνα (κλίμακα 0), κατακερματίζουμε την εικόνα σε μικρότερες κλίμακες (2×2, 4×4 κοκ), και υπολογίζουμε τον περιγραφητή σε αυτές τις μικρότερες κλίμακες, όπως φαίνεται και στο σχήμα 3.4. Ο τελικός πυραμιδωτός περιγραφητής, προκύπτει από τη συνένωση των επιμέρους περιγραφητών σε όλες τις κλίμακες, με κατάλληλη στάθμιση, ανάλογα με την κλίμακα από την οποία προέρχεται. Με αυτό τον τρόπο έχουμε εκμεταλλευτεί τη χωρική συσχέτιση που υπάρχει στην εικόνα, καθώς και την πληροφορία που υπάρχει σε διαφορετικές κλίμακες (για παράδειγμα, αντικείμενα με διαφορετικό μέγεθος). Παρά την μεγάλη περιγραφική ικανότητα των Pyramidal HOG χαρακτηριστικών, ένα σημαντικό μειονέκτημα τους είναι η μεγάλη αύξηση της διάστασης του διανύσματος χαρακτηριστικών, καθώς συνεχίζουμε τη διαδικασία εξαγωγής σε όλο και μεγαλύτερη κλίμακα.

Fourier Descriptors

Οι περιγραφητές Fourier βασίζονται στην εξαγωγή συντελεστών Fourier από το περίγραμμα του σχήματος προς περιγραφή. Στα πλαίσια της παρούσας



Σχήμα 3.4: Οπτικοποίηση της διαδικασίας εξαγωγής ιστογραφικών περιγραφητών σε διαφορετικές κλίμακες της εικόνας. Η συνένωση των περιγραφητών από τις επιμέρους περιοχές του κάθε επιπέδου, αλλά και από όλα τα επίπεδα, θα οδηγήσει στον τελικό πυραμιδωτό περιγραφητή.

διπλωματικής εργασίας, χρησιμοποιούνται σε αντιστοιχία με την εργασία των Conseil et al. [18] που χρησιμοποίησαν επίσης τους περιγραφητές Fourier για την εξαγωγή χαρακτηριστικών επί της χειρομορφής, πραγματοποιώντας τις κατάλληλες κανονικοποιήσεις, για να επιτύχουν το αναλλοίωτο των περιγραφητών στη μετατόπιση, την κλιμάκωση και την περιστροφή. Πιο συγκεκριμένα, το πρώτο βήμα πριν τον υπολογισμό των συντελεστών Fourier, είναι η δειγματοληψία του περιγράμματος, ώστε να προκύψει μια ομαλοποίηση του σχήματος, σε βαθμό τέτοιο, ώστε να κρατήσουμε βέβαια τις απαραίτητες λεπτομέρειες που μας είναι χρήσιμες. Στη συνέχεια εφαρμόζουμε το μετασχηματισμό Fourier μήκους N, που υπολογίζει τους N συντελεστές Fourier C_k. Για να προκύψουν τώρα οι περιγραφητές Fourier με τις κατάλληλες ιδιότητες, προχωράμε στις επόμενες κανονικοποιήσεις:

- Απορρίπτουμε τον πρώτο συντελεστή C_0, ο οποίος περιλαμβάνει πληροφορία μόνο για τη θέση του σχήματος.
- Αγνοούμε τη φάση, με αποτέλεσμα κάθε συντελεστής να χαρακτηρίζεται μόνο από το μέτρο του, ώστε αφενός οι περιγραφητές να είναι αναλλοίωτοι ως προς την περιστροφή, αφετέρου να υπάρχει ανεξαρτησία από το σημείο του περιγράμματος που λαμβάνεται ως αρχικό σημείο του σήματος.
- Διαιρούμε όλους τους συντελεστές ως προς το μέτρο του δεύτερου συντελεστή

C_1, ώστε να πετύχουμε ανεξαρτησία ως προς την κλίμακα.

Συνεπώς, για μετασχηματισμό Fourier μήκους N, προέκυψε ένα σύνολο $N-2$ περιγραφητών Fourier (ο πρώτος απορρίφθηκε, και ο δεύτερος προέκυψε ίσος με τη μονάδα). Μία επιπλέον μείωση της διαστασιμότητας μπορεί να επιτευχθεί με την επιλογή των N_F πρώτων τέτοιων συντελεστών, οι οποίοι περιέχουν και το μεγαλύτερο κομμάτι της χρήσιμης πληροφορίας.

Hu Moments

Οι ροπές Hu (Hu Moments) [34], αποτελούνται από επτά τιμές, που υπολογίζονται με βάση την περιοχή (region) ενός σχήματος, σε αντίθεση με τους περιγραφητές Fourier που βασίζονται στο περίγραμμα του σχήματος, και προσφέρουν μια αναπαράσταση που είναι αναλλοίωτη ως προς την κλίμακα, την περιστροφή, και τη θέση. Οι πρώτες έξι τιμές κωδικοποιούν το σχήμα, και παραμένουν αναλλοίωτες σε μετασχηματισμούς παράλληλης μεταφοράς, αλλαγής κλίμακας, και περιστροφής, ενώ η έβδομη τιμή επιτρέπει το διαχωρισμό μεταξύ μιας εικόνας, και της κατοπτρικής της.

Pattern Spectrum

Το Pattern Spectrum είναι ένας περιγραφητής που προτάθηκε από τον Maragos [48] και πρόκειται για ένα ιστόγραμμα αναπαράστασης σχήματος σε πολλές κλίμακες. Εδώ χρησιμοποιούμε την εκδοχή για δυαδικές εικόνες. Για μια δυαδική εικόνα X συνεπώς, και σε σχέση με ένα κυρτό δυαδικό δομικό στοιχείο B, το ιστόγραμμα Pattern Spectrum προκύπτει ως εξής:

$$PS_X(r,B) = \frac{-dA(X \circ rB)}{dr}, r \geq 0 \tag{3.3}$$

όπου $A(X)$ είναι το εμβαδόν της εικόνας X, ενώ η παράμετρος r ορίζει την κλίμακα. Σε αυτή την περίπτωση, το πολυκλιμακωτό opening της εικόνας X από το δομικό στοιχείο B ορίζεται ως:

$$X \circ rB = (X \ominus rB) \oplus rB = [(X \underbrace{\ominus B) \ominus B \ldots \ominus B]}_{r \text{ times}} \underbrace{\oplus B \oplus B \ldots \oplus B}_{r \text{ times}} \tag{3.4}$$

Μάλιστα με τη χρήση closings, είναι δυνατό να ορίσουμε το Pattern Spectrum και για αρνητικές τιμές του r:

$$PS_X(-r,B) = \frac{dA(X \bullet rB)}{dr}, r > 0 \tag{3.5}$$

όπου σε αντιστοιχία με το opening, το πολυκλιμακωτό closing ορίζεται ως:

$$X \bullet rB = (X \oplus rB) \ominus rB \tag{3.6}$$

HOF

Τα ιστογράμματα οπτικής ροής (Histograms of Optical Flow ή HOF), προτάθηκαν από τους Dalal et al. [22], με στόχο πάλι το πρόβλημα της αναγνώρισης ανθρώπων, αυτή τη φορά όμως σε ακολουθίες βίντεο, αντί για στατικές εικόνες. Τα HOF φιλοδοξούν να συλλάβουν τις τοπικές κατευθύνσεις των ακμών της κίνησης, και στη φιλοσοφία τους παρουσιάζουν πολλές ομοιότητες με τα HOG. Πιο συγκεκριμένα, οι συγγραφείς υπολογίζουν την οπτική ροή που προκύπτει από δύο συνεχόμενα καρέ, και τη διακρίνουν στην οριζόντια και στην κάθετη συνιστώσα της (I^x και I^y αντίστοιχα). Αντιμετωπίζοντας τις I^x και I^y ως κοινές γκρίζες εικόνες, για κάθε μία υπολογίζουν το μέτρο και την κατεύθυνση των gradients της, από τα οποία προκύπτει το ιστόγραμμα προσανατολισμού, με τρόπο αντίστοιχο με αυτό των HOG. Και πάλι, δεν εφαρμόζουν αρχικά κάποιο φίλτρο ομαλοποίησης των εικόνων, ενώ προτείνουν την πιο απλή μονοδιάστατη μάσκα, ($[-1, 0, 1]$), για τον υπολογισμό των gradients.

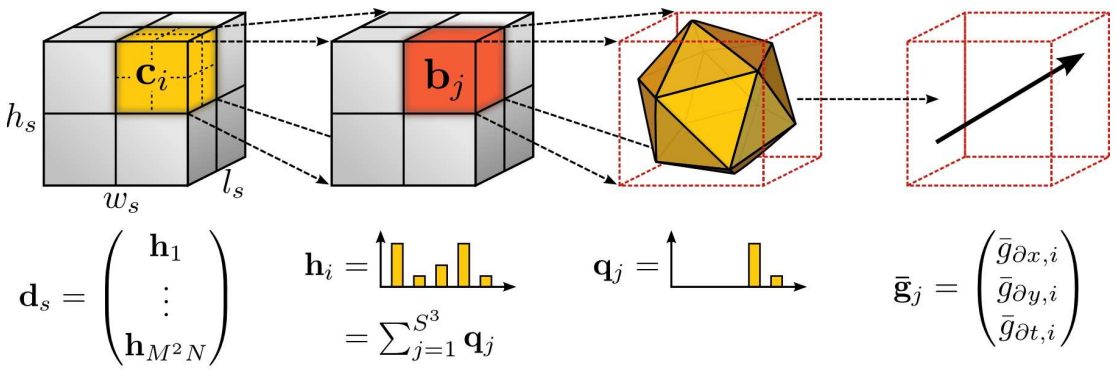
HOG/HOF

Η εισαγωγή των χαρακτηριστικών HOG/HOF, έγινε από τους Laptev et al. [43] και χρησιμοποιήθηκαν αρχικά στο πρόβλημα της αναγνώρισης δράσεων σε βίντεο. Στην ουσία πρόκειται για έναν συνδυασμό των HOG και HOF χαρακτηριστικών, με σκοπό να ενσωματωθεί στον περιγραφητή πληροφορία τόσο από την τοπική εμφάνιση, όσο και από την κίνηση. Για τον υπολογισμό του HOG/HOF περιγραφητή, η περιοχή από την οποία πρόκειται να γίνει η εξαγωγή των χαρακτηριστικών, αντιμετωπίζεται σαν ένας χωροχρονικός όγκος που είναι χωρισμένος σε ένα $n_x \times n_y \times n_t$ πλέγμα από κελιά. Σε κάθε ένα από αυτά τα κελιά γίνεται ο υπολογισμός των HOG και HOF χαρακτηριστικών, τα οποία συνενώνονται σε ένα ενιαίο διάνυσμα. Από τον συνδυασμό των περιγραφητών σε όλα τα κελιά, προκύπτει και ο τελικός HOG/HOF περιγραφητής.

HOG3D

Ο περιγραφητής HOG3D προτάθηκε από τους Kläser et al. [39] και βασίζεται σε ιστογράμματα τρισδιάστατων προσανατολισμένων gradients, ενώ μπορεί να θεωρηθεί ως μια επέκταση του κλασικού περιγραφητή SIFT στον 3D χώρο. Αφού ολοκληρωθεί ο υπολογισμός των gradients, ακολουθεί η ομοιόμορφη κβάντισή τους με βάση τον προσανατολισμό τους, χρησιμοποιώντας κανονικά πολύεδρα. Με αυτό τον τρόπο, ο περιγραφητής περιέχει πληροφορία τόσο για το σχήμα, όσο και για την κίνηση ταυτόχρονα. Όπως και στα HOG/HOF χαρακτηριστικά, κάθε χωροχρονικό τεμάχιο χωρίζεται σε $n_x \times n_y \times n_t$ κελιά. Τα επιμέρους ιστογράμματα των gradients για κάθε κελί συνενώνονται στο ίδιο διάνυσμα, το οποίο μετά από κανονικοποίηση οδηγεί στον τελικό περιγραφητή. Μία

αντιπροσωπευτική οπτικοποίηση του συγκεκριμένου περιγραφητή, δίνεται στο σχήμα 3.5 (προσαρμογή από την αρχική δημοσίευση, [39])



Σχήμα 3.5: Οπτικοποίηση της διαδικασίας εξαγωγής του περιγραφητή HOG3D, ξεκινώντας από το διαχωρισμό σε κελιά, μέχρι και την κβάντιση των 3D gradients για τον υπολογισμό των ιστογραμμάτων (προσαρμογή από [39])

3.4 Πειραματικά αποτελέσματα

Σε αυτή την ενότητα θα κάνουμε μία σύνοψη των κυριότερων πειραματικών αποτελεσμάτων με χρήση χαρακτηριστικών που έχουν εξαχθεί στη χειρομορφή. Η έρευνα έχει επεκταθεί και στις τρεις βάσεις που παρουσιάστηκαν στην υποενότητα 1.5.1, δηλαδή τη βάση στατικών χειρομορφών, την πολυτροπική βάση χειρονομιών ChaLearn, και την πολυτροπική βάση χειρονομιών MOBOT. Για την αξιολόγηση, χρησιμοποιούμε, ανάλογα με την περίπτωση, τρία διαφορετικά σχήματα εκπαίδευσης-αξιολόγησης του συστήματός μας:

- Σαφής ύπαρξη Training και Test Set. Αυτό συμβαίνει μόνο στη βάση χειρονομιών ChaLearn, η οποία έχει εκδοθεί επίσημα. Σε αυτή την περίπτωση, εφόσον υπάρχουν συγκεκριμένα σύνολα Εκπαίδευσης, Επικύρωσης, και Αξιολόγησης, χρησιμοποιούμε αυτές τις διαμερίσεις για τους πειραματισμούς μας.

- Διαμέριση 60%-40% επί του συνόλου των δεδομένων. Σε αυτή την περίπτωση κάνουμε μια τυχαία διαμέριση του συνόλου των δεδομένων σε σύνολα εκπαίδευσης και αξιολόγησης, με αναλογία 60%-40% αντίστοιχα. Προκειμένου να προκύψουν αποτελέσματα που δεν εξαρτώνται από αυτή την τυχαία διαμέριση, επαναλαμβάνουμε τη διαδικασία παραπάνω από μία φορές, και λαμβάνουμε το μέσο όρο όλων των επαναλήψεων.

- User independent πειραματισμοί. Σε αυτές τις περιπτώσεις, χρησιμοποιούμε τα δεδομένα ενός χρήστη για την αξιολόγηση του συστήματος, ενώ τα δεδομένα όλων των υπολοίπων χρηστών χρησιμοποιούνται για την εκπαίδευση. Στόχος είναι, το σύστημα να μην έχει εκπαιδευτεί με δεδομένα του χρήστη που αξιολογεί. Για το λόγο αυτό, η συγκεκριμένη περίπτωση αξιολόγησης ονομάζεται και τύπου unseen signer ή "κρυφού" χρήστη, και είναι προφανώς πιο απαιτητική από τις υπόλοιπες, εφόσον το σύστημα θα πρέπει να επιτύχει ικανοποιητική γενίκευση, ώστε να ταξινομήσει επιτυχώς δεδομένα διαφορετικής μορφής από αυτά με τα οποία έχει εκπαιδευτεί.

Στο σχολιασμό των πειραμάτων που ακολουθούν, φροντίζουμε να δηλώνεται ρητά και η μεθοδολογία που ακολουθήθηκε για την αξιολόγηση του συστήματος.

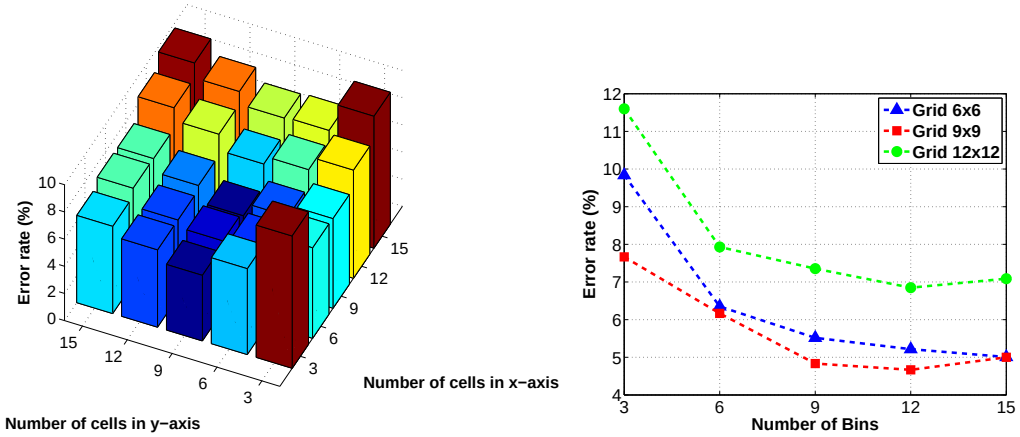
3.4.1 Πειράματα στη βάση Στατικών Χειρομορφών

Όπως περιγράψαμε και νωρίτερα, στην υποενότητα 1.5.1, για κάποια προκαταρκτικά πειράματα χρησιμοποιήσαμε μια βάση στατικών χειρομορφών, που περιλαμβάνει τα γράμματα της ελληνικής αλφαβήτου, έτσι όπως απεικονίζονται στην ελληνική νοηματική γλώσσα. Η βάση αυτή αποτέλεσε την πρώτη προσπάθεια στην κατεύθυνση ενός συστήματος αναγνώρισης χειρονομιών, πριν προσθέσουμε ακόμα τη διάστασή του χρόνου στην έρευνά μας. Κύριο εργαλείο των πειραματισμών μας εδώ κατείχε ο περιγραφητής HOG, τον οποίο εξετάσαμε αναλυτικά στα πλαίσια της συγκεκριμένης βάσης. Στη συνέχεια, θα παρουσιάσουμε κάποια αποτελέσματα από τους ευρύτερους πειραματισμούς που έγιναν [16], στο πλαίσιο που αυτά έχουν ενδιαφέρον και για την υπόλοιπη έρευνα, που εστίασε περισσότερο σε χειρονομίες παρά στατικές χειρομορφές. Σημειώνουμε ότι όλα τα πειράματα αφορούν προφανώς ταξινόμηση, για τους σκοπούς της οποίας έγινε χρήση ενός ταξινομητή τύπου kNN, με χρήση $k = 1$.

Επίδραση παραμέτρων του HOG περιγραφητή

Εφόσον για τη συγκεκριμένη βάση εστιάσαμε στη χρήση του περιγραφητή HOG, είχαμε την ευκαιρία να προχωρήσουμε και σε πιο αναλυτική εξέταση της επίδοσής του. Συγκεκριμένα, εξετάστηκε η επιρροή που έχουν οι διαφορετικές παραμετροποιήσεις του περιγραφητή (όπως το μέγεθος και το πλέγμα των κελιών, ή το πλήθος των bins), στην επίδοσή του, με σκοπό να αποκτήσουμε την απαραίτητη διαίσθηση και εμπειρία με τη χρήση του. Στο σχήμα 3.6 απεικονίζουμε τα αποτελέσματα σε αυτές τις περιπτώσεις, χρησιμοποιώντας διαμέριση των δεδομένων τύπου 60%-40% για την εκπαίδευση και αξιολόγηση του συστήματός μας.

Τα γραφήματα αυτά δίνουν μια ευρεία οπτική της επίδρασης που ασκούν οι διάφορες παράμετροι στην επίδοση του HOG περιγραφητή. Αυτό που έχει



(α') Μεταβολή πλέγματος κελιών (β') Μεταβολή πλήθους bins

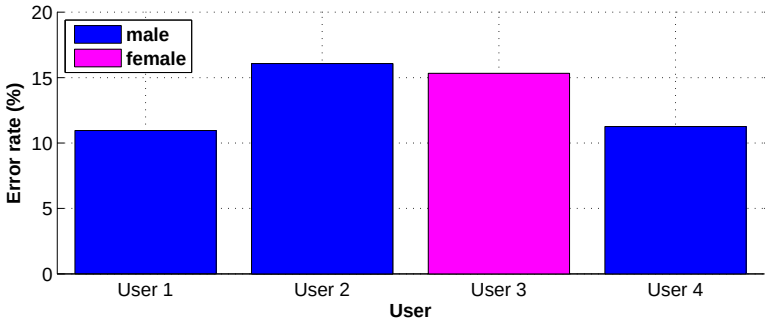
Σχήμα 3.6: Ποσοστά εσφαλμένης ταξινόμησης για διαφορετικές παραμετροποιήσεις του περιγραφητή HOG. Τα αποτελέσματα αφορούν τη βάση στατικών χειρομορφών. Αριστερά: Η επίδραση του πλέγματος των κελιών (με σταθερό αριθμό bins ίσο με 9)· Δεξιά: Η επίδραση του πλήθους των bins για σταθερό πλέγμα κελιών.

ενδιαφέρον να σχολιάσουμε εδώ είναι η σημασία της σωστής ρύθμισης των τιμών αυτών των παραμέτρων. Τόσο για το μέγεθος των κελιών, όσο και για το πλήθος των bins, παρατηρείται μια έντονη ευαισθησία σε διαφορετικές παραμετροποιήσεις. Για παράδειγμα, ένα μικρό πλέγμα κελιών, όπως 3×3, αδυνατεί να συλλάβει όλη την απαραίτητη πληροφορία της εικόνας. Από την άλλη, ένα μεγάλο πλέγμα 12×12 οδηγεί σε περιγραφητή μεγάλης διαστασιμότητας, επηρεάζοντας την επίδοση του ταξινομητή μας. Αυτές οι παρατηρήσεις θα επηρεάσουν και την υπόλοιπη έρευνα με χρήση του HOG περιγραφητή, όπου και θα απαιτήσουμε μία καλή στάθμιση των παραμέτρων για επίτευξη υψηλών ποσοστών επιτυχίας.

Πειράματα με "κρυφό" χρήστη (unseen signer)

Εφόσον τα αρχικά ποσοστά ταξινόμησης σημείωσαν αρκετά υψηλές επιδόσεις (ακόμα και υψηλότερες από 95%) για τη συγκεκριμένη βάση, επιλέξαμε να αντιμετωπίσουμε και προβλήματα "κρυφού χρήστη" ή unseen signer, που συνήθως παρουσιάζουν αυξημένη δυσκολία. Σε αυτές τις περιπτώσεις χρησιμοποιούμε τα δεδομένα ενός χρήστη για την αξιολόγηση του συστήματος, ενώ τα δεδομένα των υπολοίπων τριών χρηστών χρησιμοποιήθηκαν για την εκπαίδευση, όπως περιγράψαμε και προηγουμένως. Για τους τέσσερις χρήστες της συγκεκριμένης βάσης, λάβαμε τα αποτελέσματα που απεικονίζονται στο σχήμα 3.7

Όπως βλέπουμε εδώ, τα ποσοστά επιτυχίας αν και παραμένουν υψηλά (από



Σχήμα 3.7: Ποσοστά εσφαλμένης ταξινόμησης για πειραματισμούς τύπου unseen signer. Τα αποτελέσματα αφορούν τους χρήστες της βάσης στατικών χειρονομιών.

84% έως 89%), είναι μειωμένα σε σχέση με τους πειραματισμούς που το σύστημα είχε εκπαιδευτεί και με δεδομένα του χρήστη που αξιολογούσε. Αυτό είναι και το αναμενόμενο φυσικά, λόγω της σαφέστερης επιπλέον δυσκολίας που εισάγει το πρόβλημα τύπου unseen signer.

Σύνοψη

Σαν σύνοψη των πειραματισμών στη βάση στατικών χειρομορφών, παρουσιάζουμε τον πίνακα 3.1, που συγκεντρώνει κάποια από τα καλύτερα αποτελέσματα. Γενικά, οι επιδόσεις στη συγκεκριμένη βάση δεδομένων είναι ιδιαιτέρως υψηλές, ωστόσο πρόκειται για ένα αρκετά ευκολότερο πρόβλημα από αυτό της αναγνώρισης χειρονομιών όπου εμπλέκεται και η χρονική διάσταση.

Data	**Feat.**	**Classifier**	**Exp. Type**	**Result**	**Comments**
Static Handshapes	HOG	kNN	60%-40%	max 95.87%	Μεταβολή πλέγματος και αριθμού bins.
Static Handshapes	HOG	kNN	Unseen	83.93% έως 89.04%	9×9 cells, 9 bins.

Πίνακας 3.1: Συνοπτική παρουσίαση των αποτελεσμάτων στη βάση στατικών χειρομορφών.

3.4.2 Πειράματα στη βάση Χειρονομιών ChaLearn

Στη συνέχεια παρουσιάζουμε τα πειράματα που έγιναν στη βάση χειρονομιών ChaLearn. Η συγκεκριμένη βάση, λόγω των διαφόρων ερευνητικών προκλήσεων

που παρουσιάζει, και οι οποίες παρουσιάστηκαν στην αντίστοιχη υποενότητα 1.5.2, είναι αυτή που μας απασχόλησε περισσότερο, με αποτέλεσμα να προχωρήσει σε μεγαλύτερη έκταση ο πειραματισμός. Το κύριο ενδιαφέρον μας στηρίχτηκε στη χρήση HMMs, ωστόσο επεκταθήκαμε και σε πειράματα με άλλους δημοφιλείς ταξινομητές, όπως τα SVM και kNN, οι οποίοι όμως σε αντίθεση με τα HMMs δεν λαμβάνουν υπόψη τους τη χρονική μοντελοποίηση.

Επίσης, όλα τα πειράματα που παρουσιάζονται σε αυτή την υποενότητα αφορούν αποκλειστικά το πρόβλημα της ταξινόμησης συγκεκριμένων οπτικών αποσπασμάτων σε κάποια από τις χειρονομίες του λεξιλογίου, χωρίς να επεκταθούμε σε προβλήματα συνεχούς αναγνώρισης. Επιπλέον, έχουμε χρησιμοποιήσει το Training Set της βάσης για την εκπαίδευση του συστήματός μας, και το Validation Set για την αξιολόγηση (έτσι όπως αυτά ορίζονται από τη βάση), για τα οποία παρέχονται τα χρονικά όρια των χειρονομιών που εκτελούνται (και επομένως είναι δυνατόν να προχωρήσουμε σε ταξινόμηση αυτών).

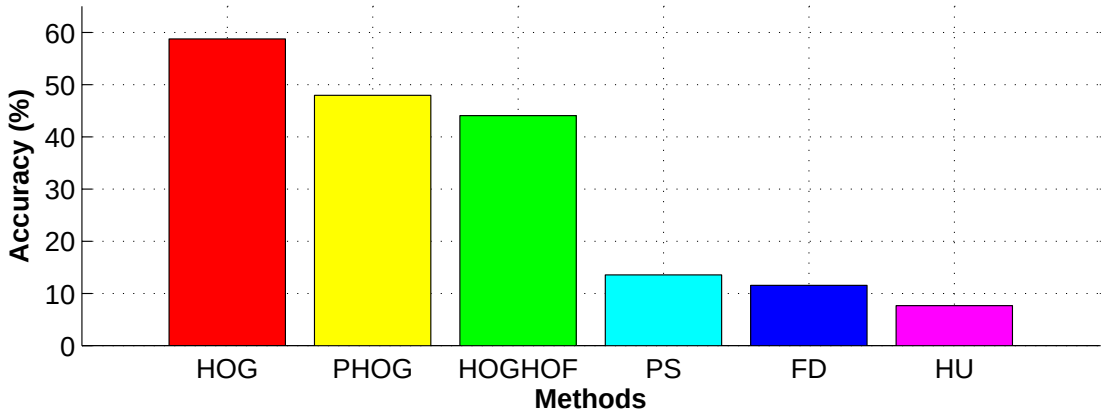
Σύγκριση διαφορετικών οπτικών περιγραφητών

Το πρώτο πείραμα που παρουσιάζουμε έχει να κάνει με τη χρήση διαφορετικών περιγραφητών για την εξαγωγή χαρακτηριστικών. Σε αυτή την περίπτωση περιοριζόμαστε στα HOG, Pyramidal HOG, HOG/HOF, Hu Moments, Pattern Spectrum και Fourier Descriptors. Η προεπεξεργασία που κάνουμε είναι ίδια με αυτή που περιγράφηκε προηγουμένως (ενότητα 3.2), για τον εντοπισμό και την κατάτμηση της χειρομορφής. Για τους περιγραφητές που επιδρούν πάνω στην πληροφορία της εμφάνισης (δηλαδή τους τρεις πρώτους από τους προαναφερόμενους), πραγματοποιήσαμε εξαγωγή χαρακτηριστικών τόσο στην πληροφορία εμφάνισης, όσο και στην πληροφορία βάθους, και το διάνυσμα χαρακτηριστικών προέκυψε ως η συνένωση των δύο επιμέρους περιγραφητών. Από την άλλη, για τους περιγραφητές που επιδρούν στο σχήμα (οι υπόλοιποι τρεις από τους προαναφερόμενους), χρησιμοποιήσαμε το σχήμα της χειρομορφής για την εξαγωγή χαρακτηριστικών. Σε κάθε περίπτωση, το διάνυσμα χαρακτηριστικών που προέκυψε, δόθηκε ως είσοδος στους HMM ταξινομητές, αφενός για εκπαίδευση, αφετέρου για αξιολόγηση.

Οι παράμετροι του κάθε περιγραφητή (για παράδειγμα ο αριθμός στις κλίμακες των Pyramidal HOG, ή το μέγεθος των κελιών και των παραθύρων στα HOG), προέκυψαν μετά από βελτιστοποίηση της επίδοσης τους σε κάθε περίπτωση. Από την άλλη, ο αριθμός των καταστάσεων των κρυφών Μαρκοβιανών μοντέλων έμεινε σταθερός στις 13 σε κάθε περίπτωση, ενώ χρησιμοποιήθηκε μία Γκαουσιανή σε κάθε κατάσταση.

Με βάση τα παραπάνω, στο Σχήμα 3.8 μπορούμε να παρατηρήσουμε τα συγκεντρωτικά αποτελέσματα και τη σύγκριση μεταξύ των διαφόρων περιγραφητών. Όπως βλέπουμε, τα ιστογραφικά χαρακτηριστικά HOG σημειώνουν το υψηλότερο

ποσοστό επιτυχημένης ταξινόμησης στη συγκεκριμένη βάση, ενώ ακολουθούν αντίστοιχες μέθοδοι που βασίζονται στον υπολογισμό gradients, όπως τα Pyramidal HOG και ο συνδυασμός HOG/HOF. Την υψηλότερη επίδοση των HOG σε σχέση με τις άλλες δύο αυτές μεθόδους, που θεωρητικά ενσωματώνουν περισσότερη πληροφορία (τα PHOG πληροφορία σε περισσότερες κλίμακες, και τα HOG/HOF πληροφορία από την οπτική ροή), την αποδίδουμε στην "ευελιξία" της παραμετροποίησης των HOG χαρακτηριστικών, που επιτρέπει την αρκετά καλή στάθμιση των παραμέτρων. Αντίθετα, τα PHOG οδηγούν σε πολύ μεγάλη αύξηση της διαστασιμότητας με τη χρήση πολλών κλιμάκων, δημιουργώντας δυσκολίες εκπαίδευσης στους ταξινομητές μας, ενώ τα HOG/HOF, έτσι όπως υλοποιούνται από τους Laptev et al. [43] δίνουν λιγότερες επιλογές παραμετροποίησης (σε χωρικές και χρονικές κλίμακες). Όσον αφορά τέλος τις υπόλοιπες μεθόδους (PS,FD,HU), αυτές σημειώνουν αρκετά χαμηλότερα ποσοστά επιτυχίας. Αυτό βέβαια είναι εν μέρει δικαιολογημένο μιας και οι Hu Moments, το Pattern Spectrum αλλά και οι Fourier Descriptors επιδρούν μόνο πάνω στο σχήμα της χειρομορφής και δεν εκμεταλλεύονται την πληροφορία εμφάνισης. Συνεπώς είναι αναμενόμενο η επίδοσή τους να είναι χαμηλότερη, λόγω της περιορισμένης πληροφορίας που αξιοποιούν.

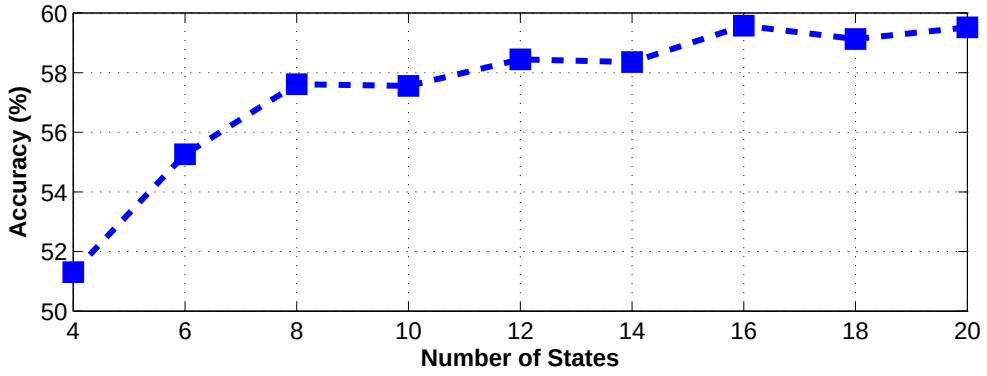


Σχήμα 3.8: Ποσοστό επιτυχημένης ταξινόμησης για χρήση διαφορετικών οπτικών περιγραφητών επί των χειρομορφών του χρήστη. Τα αποτελέσματα έχουν προκύψει από χρήση της βάσης ChaLearn.

Λαμβάνοντας υπόψη την υψηλότερη επίδοση που παρουσιάζουν τα HOG χαρακτηριστικά, στη συνέχεια, εστιάζουμε τους πειραματισμούς μας στους εν λόγω ιστογραφικούς περιγραφητές, όσον αφορά τουλάχιστον τα χαρακτηριστικά χειρομορφής.

Επίδραση του αριθμού των καταστάσεων των HMMs

Κατά τη χρήση κρυφών Μαρκοβιανών μοντέλων, έχει ιδιαίτερο ενδιαφέρον να αξιολογηθεί η επίδραση του αριθμού καταστάσεων των μοντέλων μας, αλλά και του πλήθους του Γκαουσιανών κατανομών που χρησιμοποιούμε για να περιγράψουμε τα χαρακτηριστικά κάθε κατάστασης. Στην περίπτωση των χαρακτηριστικών χειρομορφής ωστόσο, το μέγεθος του διανύσματος χαρακτηριστικών φτάνει συχνά σε πολύ μεγάλες τιμές (ακόμα και μεγαλύτερες του 500). Για το λόγο αυτό, επειδή θα έπρεπε να εκτιμηθεί πολύ μεγάλος αριθμός παραμέτρων, με συνέπεια να δημιουργηθούν προβλήματα κατά τη διαδικασία της εκπαίδευσης, περιορίσαμε των αριθμό των Γκαουσιανών κατανομών σε μία ανά κατάσταση. Εντούτοις, η αξιολόγηση της επίδρασης του αριθμού των καταστάσεων παραμένει ενδιαφέρουσα, και τα αποτελέσματά της εμφανίζονται στο σχήμα 3.9.



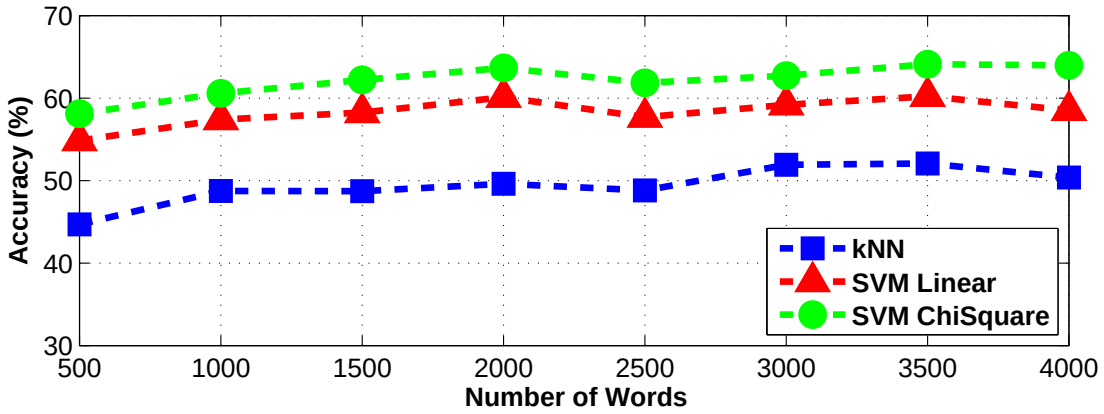
Σχήμα 3.9: Ποσοστά επιτυχημένης ταξινόμησης για διαφορετικό αριθμό καταστάσεων στα HMMs που χρησιμοποιούμε. Έχει γίνει χρήση HOG χαρακτηριστικών, ενώ τα αποτελέσματα αφορούν τη βάση ChaLearn.

Εδώ έχουμε παρουσιάσει τα ποσοστά επιτυχημένης ταξινόμησης για χρήση μοντέλων με διαφορετικό αριθμό καταστάσεων, αρχίζοντας από 4 και φτάνοντας μέχρι και τις 20 καταστάσεις. Αυτό που παρατηρούμε είναι ότι γενικά φαίνεται να υπάρχει αύξηση των ποσοστών επιτυχίας καθώς αυξάνεται και ο αριθμός των καταστάσεων. Αυτή η αυξητική τάση εξαλείφεται για περισσότερες από 15 καταστάσεις, οπότε και το ποσοστό ακρίβειας της ταξινόμησης παραμένει εν γένει σταθερό. Αξίζει να σημειώσουμε ότι η μέγιστη επίδοση επιτυχημένης ταξινόμησης, σημειώνεται για 16 καταστάσεις στα μοντέλα μας, και ξεπερνάει το *59%*

Χρήση ‘‘στατικών’’ ταξινομητών

Ο λόγος που επικεντρωθήκαμε σε ταξινομητές τύπου HMM, είναι η πεποίθησή μας ότι αυτοί μπορούν να περιγράψουν καλύτερα τις χειρονομίες που σκοπεύουμε να

αναγνωρίσουμε, μιας και μοντελοποιούν τη χρονική εξέλιξη ενός φαινομένου, όπως για παράδειγμα η ομιλία, ή η εκτέλεση ενός νοήματος της νοηματικής γλώσσας. Ωστόσο, έχει ενδιαφέρον να κάνουμε μια σύγκριση τους με πιο "στατικούς" ταξινομητές, οι οποίοι είναι ιδιαιτέρως δημοφιλείς στη βιβλιογραφία, όπως τα Support Vector Machines ή SVMs, αλλά και οι ταξινομητές k-Nearest Neighbor ή kNN. Για το σκοπό αυτό, θα εφαρμόσουμε πρώτα την τεχνική Bag-of-Features που εξετάσαμε στην υποενότητα 2.2.3, έτσι ώστε να φέρουμε τα αποσπάσματα βίντεο σε μορφή κατάλληλη για τους ταξινομητές μας. Να σημειώσουμε, ότι η χρήση αυτής της τεχνικής δεν είναι εφικτή με το "στατικό" ανάλογο των HMMs, τα GMMs (HMMs μίας κατάστασης), εφόσον τα ιστογράμματα που παράγονται από την τεχνική Bag-of-Features είναι ιδιαίτερως αραιά-sparse, και είναι δύσκολο να εκπαιδευτούν τα μοντέλα μας. Με βάση τα παραπάνω, εκτελούμε τον πειραματισμό μας, τα αποτελέσματα του οποίου παρουσιάζονται στο σχήμα 3.10.



Σχήμα 3.10: Ποσοστά επιτυχημένης ταξινόμησης με τη χρήση της τεχνικής Bag-of-Features για χαρακτηριστικά χειρομορφής και "στατικούς" ταξινομητές (kNN και SVM με γραμμική ή χ^2 απόσταση). Τα αποτελέσματα αφορούν τη βάση χειρονομιών ChaLearn

Όπως παρατηρούμε, επιτυγχάνονται υψηλά ποσοστά επιτυχημένης ταξινόμησης, ακόμα ψηλότερα από αυτά των HMM. Συγκεκριμένα, για χρήση SVM με χ^2-απόσταση (που συγκεντρώνει τις καλύτερες επιδόσεις από τους ταξινομητές που χρησιμοποιήσαμε), το ποσοστό επιτυχημένης ταξινόμησης ξεπερνάει ακόμα και το *64%*, τη στιγμή που για τα HMMs, όπως είδαμε νωρίτερα ήταν μόλις λίγο ψηλότερα από το 59%. Το στοιχείο αυτό είναι βεβαίως θετικό, όμως δεν θα πρέπει να ξεχνάμε ότι τέτοιου είδους ταξινομητές προορίζονται αποκλειστικά για "στατικά" προβλήματα ταξινόμησης. Συνεπώς δεν είναι εμφανές, ούτε εύκολο το πως θα μπορούσαμε να επεκτείνουμε τη χρήση τους σε προβλήματα συνεχούς αναγνώρισης.

Σύνοψη

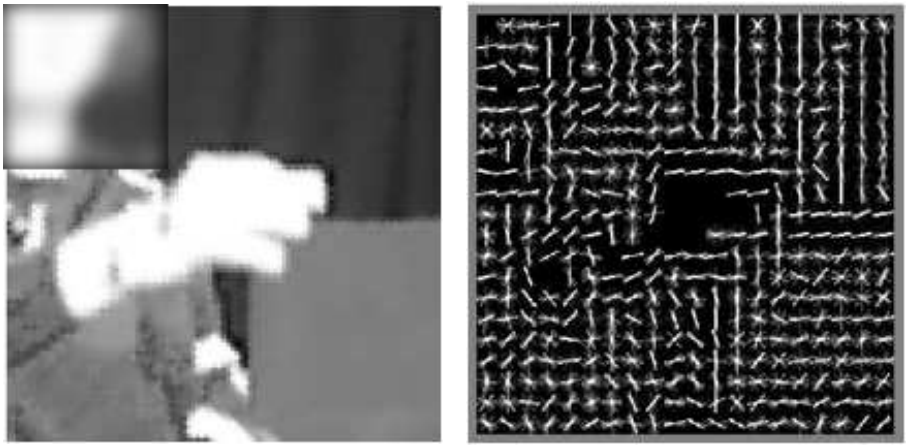
Ολοκληρώνοντας την ανάλυση των πειραματισμών στη βάση χειρονομιών ChaLearn, κάνουμε μια συγκεντρωτική παρουσίαση των αποτελεσμάτων στον πίνακα 3.2. Όπως παρατηρούμε, τα ποσοστά επιτυχίας για το πρόβλημα της ταξινόμησης ξεπερνούν το 59% με χρήση HMMs, ενώ υπερβαίνουν το 64% με χρήση SVM. Οι επιδόσεις είναι αρκετά ικανοποιητικές, ωστόσο, όπως θα δούμε στη συνέχεια (κεφάλαιο 4), με εκμετάλλευση ενός άλλου καναλιού οπτικής πληροφορίας (αυτού της Θέσης-Κίνησης), είναι δυνατόν να επιτύχουμε ακόμα υψηλότερα ποσοστά επιτυχίας.

Data	**Feat.**	**Classifier**	**Exp. Type**	**Result**	**Comments**
ChaLearn	Various	HMM	Training-Validation	max 58.74%	13 states, 1 mixture.
ChaLearn	HOG	HMM	Training-Validation	max 59.58%	4-20 states, 1 mixture.
ChaLearn	HOG+BoF	kNN/SVM	Training-Validation	max 64.13%	500-4000 visual words.

Πίνακας 3.2: Συνοπτική παρουσίαση των αποτελεσμάτων στη βάση χειρονομιών ChaLearn, χρησιμοποιώντας το κανάλι πληροφορίας της Χειρομορφής.

3.4.3 Πειράματα στη βάση Χειρονομιών MOBOT

Αντίστοιχο ενδιαφέρον παρουσιάζει και η επίσης πολυτροπική βάση χειρονομιών MOBOT, για την εκτέλεση αντίστοιχων πειραμάτων. Μάλιστα, λόγω του μικρότερου μεγέθους της, αλλά και των προβλημάτων που περιγράψαμε και προηγουμένως στην παρουσίαση της βάσης, στην υποενότητα 1.5.3, θα ήταν σημαντικός ένας ειδικός χειρισμός για την αποτελεσματικότερη αντιμετώπιση των ερευνητικών προκλήσεων που παρουσιάζει. Ωστόσο, επειδή αυτός ο σκοπός ξεφεύγει από τα πλαίσια της παρούσας διπλωματικής, επιλέξαμε να εστιάσουμε κυρίως στο πειραματισμό με μεθόδους ή ιδέες που έχουν ήδη φανεί αποδοτικές στη βάση ChaLearn, όπου πραγματοποιήθηκε και το μεγαλύτερο εύρος πειραμάτων, ώστε να αξιολογήσουμε τη χρήση τους σε μια νέα βάση δεδομένων, με την οποία δεν είμαστε εξοικειωμένοι. Συνεπώς, τουλάχιστον για τα χαρακτηριστικά της χειρομορφής, ο πειραματισμός εστίασε στη χρήση HOG χαρακτηριστικών (παράδειγμα των οποίων δίνεται στο σχήμα 3.11), με χρήση κρυφών Μαρκοβιανών μοντέλων. Επιπλέον, εφόσον ήταν χρήσιμο να εξεταστεί και να αξιολογηθεί η έντονη διαφοροποίηση στην εκτέλεση χειρονομιών μεταξύ των διαφόρων χρηστών, η εκπαίδευση-αξιολόγηση ήταν τύπου unseen signer.



(α') Χειρομορφή (β') Περιγραφητής HOG

Σχήμα 3.11: Παράδειγμα εξαγωγής του περιγραφητή HOG σε χειρομορφή της βάσης χειρονομιών MOBOT. Αριστερά: Στιγμιότυπο της χειρομορφής κατά την εκτέλεση χειρονομίας. Δεξιά: Οπτικοποίηση του HOG περιγραφητή στη συγκεκριμένη εικόνα. Παρά τη μέτρια ποιότητα της εικόνας εμφάνισης, ο περιγραφητής HOG έχει καταφέρει να "συλλάβει" αρκετή χρήσιμη πληροφορία, κυρίως για το σχήμα της χειρομορφής.

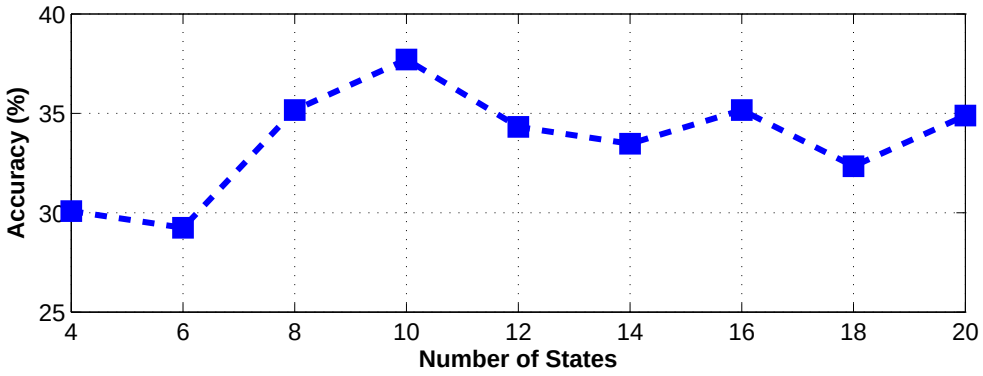
Επίδραση του αριθμού των καταστάσεων των HMMs

Και πάλι μελετήσαμε την επίδραση του διαφορετικού αριθμού καταστάσεων στα HMMs, μία από τις κύριες παραμέτρους των μοντέλων μας. Για τους λόγους που εξηγήσαμε και παραπάνω, στον αντίστοιχο πειραματισμό στη βάση ChaLearn, δεν επεκταθήκαμε σε παραπάνω από μία Γκαουσιανές κατανομές ανά κατάσταση και ανά χαρακτηριστικό. Επικεντρώνοντας έτσι στον αριθμό των καταστάσεων, παρουσιάζουμε τα συγκεντρωτικά αποτελέσματα στο σχήμα 3.12.

Τα αποτελέσματα που βλέπουμε, με μια πρώτη ματιά δεν παρουσιάζουν σταθερά ανοδική πορεία σε σχέση με τον αριθμό των καταστάσεων, όπως συνέβαινε με τη βάση ChaLearn. Αντίθετα, το μέγιστο ποσοστό παρατηρείται στις 10 καταστάσεις, με τα μοντέλα με κοντινό πλήθος καταστάσεων να παρουσιάζουν αντίστοιχα ποσοστά επιτυχίας.

Αξιολόγηση επίδοσης ανά χρήστη

Πέρα από τα παραπάνω αποτελέσματα, που είναι συγκεντρωτικά για όλους τους χρήστες, κάνουμε και μια αναλυτική αξιολόγηση με βάση την επίδοση σε κάθε χρήστη ξεχωριστά. Η τακτική για εκπαίδευση και αξιολόγηση παραμένει η ίδια (unseen signer), ωστόσο τώρα εστιάζουμε στο ποσοστό επιτυχημένης ταξινόμησης

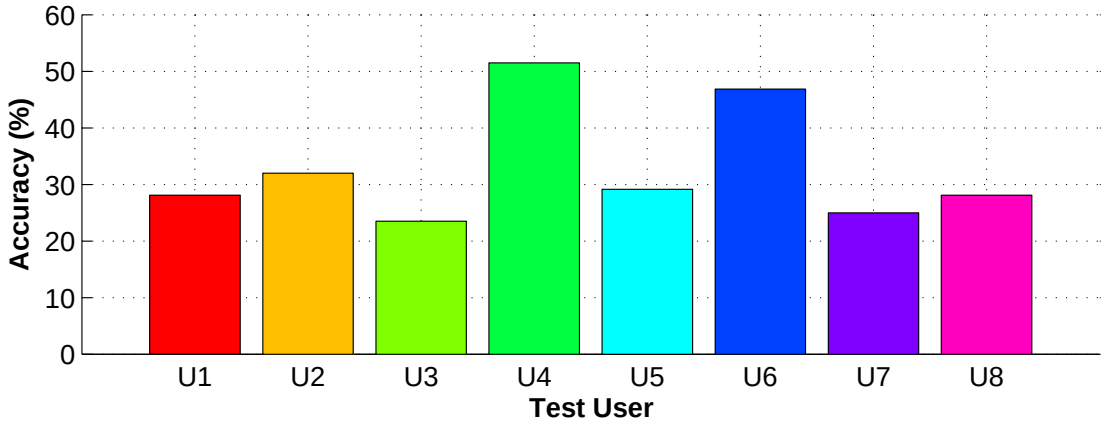


Σχήμα 3.12: Ποσοστά επιτυχημένης ταξινόμησης, για διαφορετικό αριθμό καταστάσεων στα HMMs που χρησιμοποιούμε επί της πληροφορίας της Χειρομορφής. Έγινε χρήση HOG χαρακτηριστικών, ενώ τα αποτελέσματα αφορούν τη βάση MOBOT.

κάθε χρήστη ανεξάρτητα. Όσον αφορά τον αριθμό των καταστάσεων, αυτός έχει σταθεροποιηθεί στις 10, οι οποίες είδαμε νωρίτερα ότι μεγιστοποιούν την επίδοση του συστήματος μας, ενώ και πάλι χρησιμοποιούμε HOG χαρακτηριστικά. Τα πλήρη αποτελέσματα παρουσιάζονται στο σχήμα 3.13. Όπως μπορούμε να παρατηρήσουμε, υπάρχει μία γενικότερη συμφωνία των ποσοστών επιτυχίας μεταξύ των διαφόρων χρηστών, με εξαίρεση δύο από αυτούς οι οποίοι παρουσιάζουν εμφανώς υψηλότερη επίδοση.

Σύνοψη

Εν κατακλείδι, όσον αφορά τον πειραματισμό στη βάση χειρονομιών MOBOT, παρουσιάζουμε συγκεντρωτικά τα αποτελέσματά μας στον πίνακα 3.3. Όπως παρατηρούμε, τουλάχιστον για τα χαρακτηριστικά της Χειρομορφής, τα αποτελέσματα είναι συγκριτικά μειωμένα σε σχέση με όσα παρατηρήσαμε στη βάση ChaLearn. Όπως προείπαμε, λόγω της ιδιαιτερότητας του προβλήματος θα ήταν χρήσιμη μία διαφορετική προσέγγιση που να εστιάζει στις ανάγκες και τις δυσκολίες της συγκεκριμένης περίπτωσης. Ωστόσο, είναι ικανοποιητικό, το ότι με την εφαρμογή μιας μεθόδου που εστιάζει στις ανάγκες ενός προβλήματος διαφορετικής λογικής, καταφέραμε και επιτύχαμε μία αξιόλογη επίδοση (ποσοστό επιτυχημένης ταξινόμησης μεγαλύτερο του 37%) σε μία ομολογουμένως απαιτητική τεχνική αξιολόγησης, όπως αυτή του unseen signer.



Σχήμα 3.13: Ποσοστά επιτυχημένης ταξινόμησης, για κάθε χρήστη ξεχωριστά. Σε κάθε περίπτωση, το σύστημα έχει εκπαιδευτεί με όλους τους χρήστες, πλην αυτού που χρησιμοποιείται για την αξιολόγηση (unseen signer πείραμα). Έχει γίνει χρήση HOG χαρακτηριστικών και HMM ταξινομητών, ενώ τα αποτελέσματα αφορούν τη βάση χειρονομιών MOBOT.

Data	**Feat.**	**Classifier**	**Exp. Type**	**Result**	**Comments**
MOBOT	HOG	HMM	Unseen	max 37.71%	4-20 states, 1 mixture.
MOBOT	HOG	HMM	Unseen	23.53% έως 51.52%	10 states, 1 mixture.

Πίνακας 3.3: Συνοπτική παρουσίαση των αποτελεσμάτων στη βάση χειρονομιών MOBOT, χρησιμοποιώντας το κανάλι πληροφορίας της Χειρομορφής.

Κεφάλαιο 4

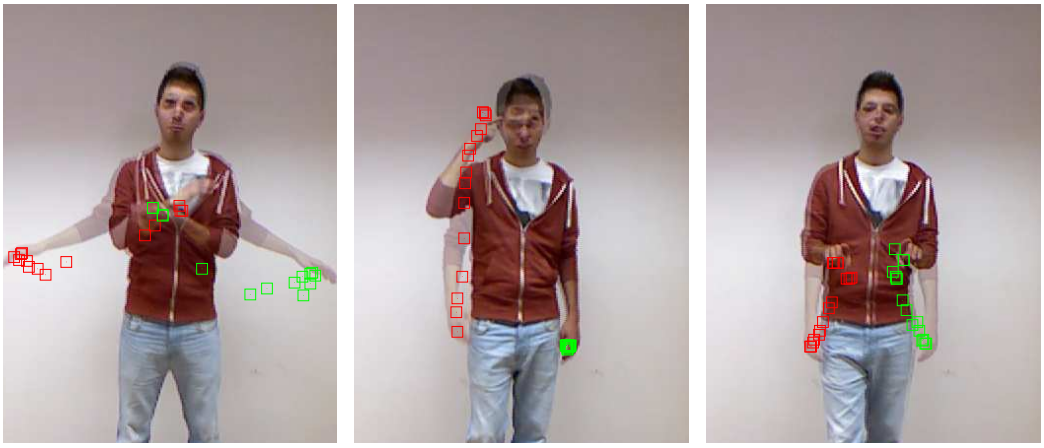
Εκμετάλλευση πληροφορίας Θέσης - Κίνησης

4.1 Γενικά

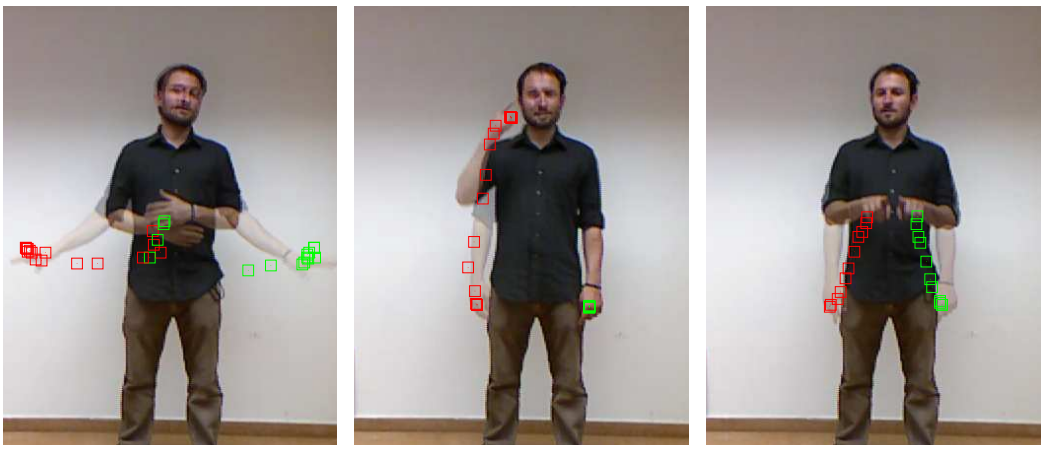
Στο προηγούμενο κεφάλαιο επικεντρωθήκαμε στη χειρομορφή του χρήστη που εκτελεί χειρονομίες, η οποία αποτέλεσε και το κύριο κανάλι πληροφορίας. Πραγματοποιήσαμε εξαγωγή χαρακτηριστικών που περιελάμβαναν πληροφορία για την εμφάνιση της χειρομορφής, το σχήμα της, ακόμα και την κίνησή της σε σχέση με το χρόνο. Ωστόσο, μία χειρονομία δεν καθορίζεται αποκλειστικά από την χειρομορφή του χρήστη. Όσο σημαντικό ρόλο και αν έχει, υπάρχουν περιπτώσεις που υπερισχύουν άλλα χαρακτηριστικά, όπως η κίνηση των χεριών, αλλά και η θέση τους κατά την εκτέλεση των χειρονομιών (ένα τέτοιο παράδειγμα παρουσιάζεται στο σχήμα 4.1). Και μπορεί κάποια πληροφορία σχετική με την κίνηση τουλάχιστον να εμπεριεχόταν ως ένα βαθμό σε περιγραφητές όπως οι HOG/HOF ή ο HOG3D, παρ' όλα αυτά ένας ακόμα πιο άμεσος τρόπος είναι να υπολογιστούν με συγκεκριμένο τρόπο τα χαρακτηριστικά της θέσης και της κίνησης των χεριών.

Και πάλι η εργασία μας στηρίζεται σε μία από τις καινοτομίες του Kinect, και πιο συγκεκριμένα στον υπολογισμό του σκελετού του χρήστη που βρίσκεται στο πεδίο λήψης του. Αυτή η πληροφορία αποδεικνύεται εξαιρετικά χρήσιμη για το πρόβλημα της αναγνώρισης χειρονομιών, αφού είναι πλέον τετριμμένος ο υπολογισμός βασικών χαρακτηριστικών της Θέσης (σχετικές θέσεις χεριών, αγκώνων κλπ) και της Κίνησης (ταχύτητες, διευθύνσεις κίνησης κλπ). Με δεδομένη αυτή την πληροφορία συνεπώς, το κεφάλαιο που ακολουθεί είναι αφιερωμένο στην αναγνώριση χειρονομιών με χρήση χαρακτηριστικών Θέσης-Κίνησης. Αρχικά παρουσιάζουμε τη μεθοδολογία της εργασίας μας, και τα χαρακτηριστικά Θέσης-Κίνησης που χρησιμοποιήσαμε, ενώ στη συνέχεια ακολουθούν τα αποτελέσματα

του πειραματισμού στη βάση ChaLearn και στη βάση MOBOT.



(α') "Basta" - Χρήστης 1 (β') "Seipazzo" - Χρήστης 1 (γ') "Daccordo" - Χρήστης 1



(δ') "Basta" - Χρήστης 2 (ε') "Seipazzo" - Χρήστης 2 (ϝ') "Daccordo" - Χρήστης 2

Σχήμα 4.1: Μεταβολή της θέσης των χεριών κατά την εκτέλεση διαφόρων χειρονομιών. Παρουσιάζονται οπτικοποιήσεις για τις χειρονομίες "basta", "seipazzo" και "daccordo" από δύο διαφορετικούς χρήστες. Οι θέσεις που φαίνεται να ακολουθούν τα χέρια σε κάθε περίπτωση παρουσιάζουν έντονες διαφοροποιήσεις για διαφορετικές χειρονομίες, ωστόσο φαίνεται να υπάρχει συσχέτιση για την εκτέλεση της ίδιας χειρονομίας από διαφορετικούς χρήστες.

4.2 Μεθοδολογία Εξαγωγής Χαρακτηριστικών

Όπως αναφέρθηκε και προηγουμένως, ο υπολογισμός των χαρακτηριστικών Θέσης-Κίνησης στην περίπτωσή μας είναι κατά κύριο λόγο τετριμμένος. Η πρόκληση στην περίπτωσή μας είναι να εντοπίσουμε τα κατάλληλα χαρακτηριστικά που θα περιγράψουν και θα μοντελοποιήσουν τη θέση και την κίνηση των χεριών του χρήστη κατά την εκτέλεση χειρονομιών. Στόχος μας είναι με την κατάλληλη μοντελοποίηση της κάθε χειρονομίας, αυτή να είναι εύκολα διακριτή από ένα σύστημα αναγνώρισης, τόσο σε σχέση με τις άλλες διαθέσιμες χειρονομίες, όσο και σε σχέση με άλλου είδους κινήσεις, οι οποίες μπορεί να βρίσκονται εκτός του λεξιλογίου, και να αφορούν είτε άλλες χειρονομίες, είτε αντανακλαστικές κινήσεις (για παράδειγμα τρίψιμο μιας περιοχής του σώματος).

Για να κάνουμε πιο συγκεκριμένο το είδος των χαρακτηριστικών Θέσης-Κίνησης, στη συνέχεια απαριθμούμε κάποια παραδείγματα τέτοιων χαρακτηριστικών που μπορούμε να χρησιμοποιήσουμε:

- 3D σχετική θέση των χεριών ως προς το κεφάλι.
- 3D σχετική θέση του δεξιού χεριού ως προς το αριστερό.
- 3D ταχύτητα κίνησης των χεριών.
- 3D διεύθυνση κίνησης των χεριών.
- 3D επιτάχυνση κίνησης των χεριών.
- 3D σχετική θέση των χεριών ως προς τους αγκώνες.
- 3D σχετική θέση των αγκώνων ως προς το κεφάλι.

Ασφαλώς η λίστα μπορεί να συνεχιστεί ακόμα περισσότερο, να προστεθεί πληροφορία θέσης-κίνησης που να αφορά κάποιο άλλο σημείο του σώματος (για παράδειγμα τους καρπούς, ή τους αγκώνες), να προστεθεί κάποια άλλη κανονικοποίηση των χαρακτηριστικών (για παράδειγμα ως προς την αρχική πόζα ηρεμίας) κοκ. Ωστόσο για τις ανάγκες του πειραματισμού μας τουλάχιστον επιλέξαμε αυτά τα χαρακτηριστικά, που κατά κύριο λόγο εμπεριέχουν το μεγαλύτερο μέρος της χρήσιμης πληροφορίας και έχει ενδιαφέρον να τα αξιολογήσουμε.

Όσον αφορά το ποια χαρακτηριστικά από τα παραπάνω θα επιλεγούν, αυτό θα αξιολογηθεί σε δύο επίπεδα:

- Αφενός μπορούμε να κάνουμε μια επιλογή των πιο σημαντικών χαρακτηριστικών που βοηθούν στη διαχωρισιμότητα μεταξύ των διαφόρων

χειρονομιών. Για παράδειγμα σε χειρονομίες που εκτελούνται με ένα από τα δύο χέρια, οι πληροφορίες για τη θέση και την κίνηση του δευτερεύοντος χεριού εισάγουν θόρυβο στο σύστημά μας, και συνεπώς είναι όχι απλά περιττές, αλλά και ανεπιθύμητες.

- Αφετέρου θα γίνει μια δεύτερη επιλογή, η οποία θα προκύψει από πειραματισμό με διάφορους συνδυασμούς των διαθέσιμων χαρακτηριστικών. Πιο συγκεκριμένα, μπορεί κάποια χαρακτηριστικά να προσφέρουν "χρήσιμη" πληροφορία στο σύστημά μας (με την έννοια ότι δεν εισάγουν θόρυβο), αλλά από τα ποσοστά αναγνώρισης να προκύπτει ότι δεν οδηγούν σε διαχωρισιμότητα μεταξύ των διαφορετικών κλάσεων χειρονομιών. Για παράδειγμα η πληροφορία της επιτάχυνσης των χεριών, πιθανώς να μη βοηθάει στη διαφοροποίηση μεταξύ χειρονομιών εφόσον υπάρχει ήδη η πληροφορία της ταχύτητας.

4.3 Πειραματικά αποτελέσματα

Για τους πειραματισμούς μας με χρήση των χαρακτηριστικών θέσης-κίνησης εστιάσαμε στις πολυτροπικές βάσεις χειρονομιών ChaLearn και MOBOT, για τις οποίες και παρουσιάζουμε αναλυτικά αποτελέσματα στη συνέχεια. Σε αντιστοιχία με την εισαγωγή που κάναμε στην ενότητα 3.4 που αφορούσε πειραματισμό με χαρακτηριστικά που έχουν προέλθει από την πληροφορία της χειρομορφής, εδώ χρησιμοποιούμε δύο σχήματα αξιολόγησης του συστήματός μας· είτε την αξιοποίηση των ήδη ορισμένων συνόλων εκπαίδευσης-επικύρωσης-αξιολόγησης για τη βάση ChaLearn, είτε τη μεθοδολογία τύπου unseen signer για τη βάση MOBOT.

4.3.1 Πειράματα στη βάση Χειρονομιών ChaLearn

Αρχικά εστιάζουμε στην πολυτροπική βάση χειρονομιών ChaLearn, και παρουσιάζουμε τους εκτεταμένους πειραματισμούς μας σε αυτή, οι οποίοι καταλαμβάνουν την κύρια έκταση αυτής της ενότητας. Και πάλι αξιοποιήσαμε κυρίως τα HMMs για τη μοντελοποίηση των χειρονομιών, με μικρή αναφορά στους ταξινομητές SVM και kNN. Όπως και για τους πειραματισμούς στη ροή πληροφορίας της χειρομορφής, παρουσιάζουμε αποκλειστικά αποτελέσματα από προβλήματα ταξινόμησης, ενώ χρησιμοποιείται το Training Set της βάσης για την εκπαίδευση, και το Validation Set για την αξιολόγηση.

Επιλογή διανύσματος χαρακτηριστικών

Ο πρώτος πειραματισμός που θα κάνουμε, σχετίζεται με την επιλογή των μεγεθών/ποσοτήτων που θα χρησιμοποιηθούν στο διάνυσμα χαρακτηριστικών.

Προηγουμένως ορίσαμε ένα πλήθος χαρακτηριστικών που έχει ενδιαφέρον να ενσωματωθούν στο διάνυσμά μας, ωστόσο είναι πιθανό να προκύψει ότι μόνο ένα συγκεκριμένο υποσύνολο από αυτά μπορεί να προσφέρει χρήσιμη πληροφορία, η οποία να βελτιώνει τις επιδόσεις αναγνώρισης.

Η τεχνική που εφαρμόζουμε στη συνέχεια, έχει να κάνει με την διαδοχική προσθήκη και αφαίρεση αυτών των ποσοτήτων στο διάνυσμα χαρακτηριστικών, και στη συνέχεια η αξιολόγησή του με βάση το ποσοστό επιτυχημένης ταξινόμησης. Τα χαρακτηριστικά μας τα διαχωρίζουμε σε τρεις κατηγορίες:

1. Σε αυτά που αφορούν κυρίως τη θέση των χεριών, δηλαδή τη σχετική θέση των χεριών ως προς το κεφάλι, και τη σχετική θέση μεταξύ τους.

2. Σε αυτά που αφορούν την κίνηση των χεριών, και πιο συγκεκριμένα την ταχύτητα, την επιτάχυνση, αλλά και τη διεύθυνση της κίνησης.

3. Σε αυτά που προσθέτουν πληροφορία σχετικά με τους αγκώνες, όπως τη σχετική θέση των αγκώνων ως προς το κεφάλι, αλλά και τη σχετική θέση των χεριών ως προς τους αγκώνες.

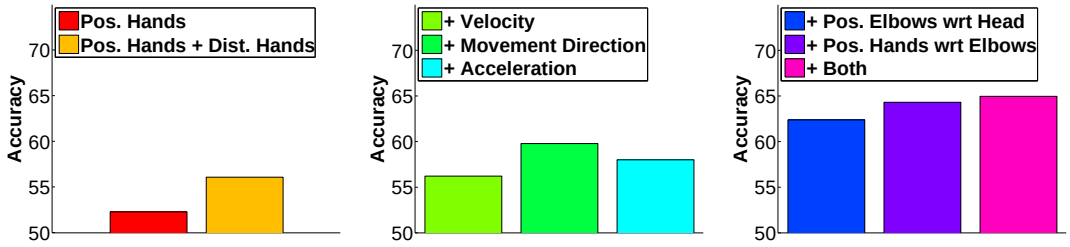
Σημειώνουμε εδώ, πριν την παρουσίαση του εκτενούς πειραματισμού, ότι σε όλες τις εκτελέσεις με διαφορετικά διανύσματα χαρακτηριστικών, κρατήσαμε σταθερές όλες τις υπόλοιπες παραμέτρους του πειράματος. Η εκπαίδευση και η αξιολόγηση έγινε με τον ίδιο τρόπο, και τα μοντέλα μας είχαν σταθερό αριθμό καταστάσεων (ίσο με 13), και από μία Γκαουσιανή ανά κατάσταση.

Προχωρώντας τώρα στους πειραματισμούς, αρχικά, από την πρώτη κατηγορία, θεωρούμε ότι τουλάχιστον η σχετική θέση των χεριών ως προς το κεφάλι είναι θεμελιώδες μέγεθος, οπότε αποτελεί τη βάση του διανύσματός μας. Έτσι πειραματιζόμαστε αντίστοιχα με την προσθήκη ή όχι της σχετικής θέσης των δύο χεριών μεταξύ τους. Όπως παρατηρούμε στο σχήμα 4.2α′, η επίδοση βελτιώνεται με την προσθήκη των χαρακτηριστικών της απόστασης των χεριών, οπότε ενσωματώνονται στο διάνυσμα χαρακτηριστικών.

Στη συνέχεια, εξετάζουμε την προσθήκη των διαφόρων χαρακτηριστικών που αφορούν την κίνηση των χεριών, και τα οποία εντάξαμε στη δεύτερη κατηγορία προηγουμένως. Έχοντας κρατήσει σταθερό το διάνυσμα με τα δύο πρώτα μεγέθη, πειραματιζόμαστε σε τρεις διαφορετικές περιπτώσεις με την επιπλέον προσθήκη είτε της ταχύτητας, είτε της επιτάχυνσης, είτε της διεύθυνσης κίνησης των χεριών. Τα τελικά αποτελέσματα παρουσιάζονται στο σχήμα 4.2β′. Όπως βλέπουμε, η τρίτη περίπτωση με χρήση της διευθύνσεως υπερέχει των υπολοίπων, οπότε η διεύθυνση κίνησης των χεριών γίνεται το επόμενο χαρακτηριστικό που προστίθεται στο διάνυσμά μας.

Τέλος, σειρά έχουν τα χαρακτηριστικά σχετικά με τους αγκώνες· η σχετική θέση των χεριών ως προς τους αγκώνες και η σχετική θέση των αγκώνων ως

προς το κεφάλι. Με σταθερό πάλι το διάνυσμα μέχρι τώρα, προσθέτουμε είτε τα χαρακτηριστικά της σχετικής θέσης χεριού-αγκώνων, είτε αυτά της σχετικής θέσης αγκώνων-κεφαλιού, είτε και τα δύο ταυτόχρονα. Τα αποτελέσματά μας παρουσιάζονται στο σχήμα 4.2γ΄. Όπως βλέπουμε με την προσθήκη και των δύο μεγεθών έχουμε την καλύτερη επίδοση, οπότε διατηρούνται και τα δύο στο διάνυσμά μας.



(α΄) Θέση Χεριών (β΄) Κίνηση Χεριών (γ΄) Θέση Αγκώνων

Σχήμα 4.2: Ποσοστό επιτυχημένης ταξινόμησης με χρήση διαφορετικών μεγεθών στο διάνυσμα των χαρακτηριστικών Θέσης-Κίνησης. Στο (α΄) παρουσιάζονται αποτελέσματα για χαρακτηριστικά που αφορούν τη θέση των χεριών. Στο (β΄) στο διάνυσμα χαρακτηριστικών έχει προστεθεί πληροφορία και σχετικά με την κίνηση των χεριών. Τέλος στο (γ΄) το διάνυσμα περιλαμβάνει και χαρακτηριστικά που αφορούν τους αγκώνες. Όλα τα αποτελέσματα αφορούν τη βάση χειρονομιών ChaLearn.

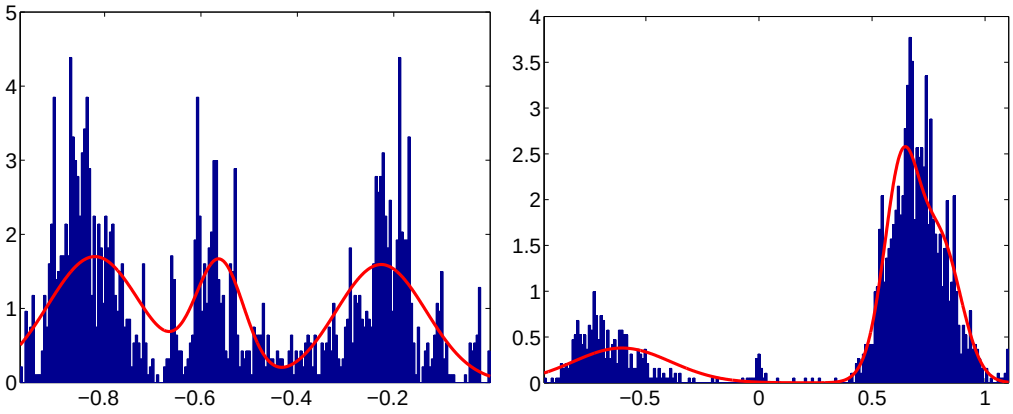
Τελικά, μετά από πειραματισμό καταλήγουμε σε ένα διάνυσμα χαρακτηριστικών που περιέχει:

- 3D σχετική θέση των χεριών ως προς το κεφάλι.
- 3D σχετική θέση του δεξιού χεριού ως προς το αριστερό.
- 3D διεύθυνση κίνησης των χεριών.
- 3D σχετική θέση των χεριών ως προς τους αγκώνες.
- 3D σχετική θέση των αγκώνων ως προς το κεφάλι.

Το συγκεκριμένο διάνυσμα είναι αυτό που χρησιμοποιήθηκε ως επί το πλείστον στους πειραματισμούς που παρουσιάζονται στη συνέχεια, εκτός ελαχίστων εξαιρέσεων, οι οποίες θα αναφέρονται όπου γίνονται.

Επίδραση αριθμού καταστάσεων και πλήθους Γκαουσιανών ανά κατάσταση

Όπως και στα χαρακτηριστικά της χειρομορφής, έτσι και εδώ επιδιώκουμε να εξετάσουμε την επίδραση που έχει ο αριθμός των καταστάσεων των μοντέλων μας. Ωστόσο, σε αυτή την περίπτωση, εφόσον προέκυψε ένα συμπαγές διάνυσμα χαρακτηριστικών διάστασης 27, είναι πιο εύκολο πλέον να αυξήσουμε το πλήθος των Γκαουσιανών που χρησιμοποιούμε για να περιγράψουμε την κατανομή κάθε χαρακτηριστικού σε κάθε κατάσταση. Μάλιστα αυτή η συγκεκριμένη τακτική πιστεύουμε ότι θα δώσει καλύτερη επίδοση στο σύστημά μας, μιας και είναι σύνηθες και αναμενόμενο η κατανομή των χαρακτηριστικών να μην περιγράφεται με το βέλτιστο τρόπο με μία Γκαουσιανή, αλλά να είναι απαραίτητο ένα μείγμα Γκαουσιανών. Παραδείγματα αυτής της περίπτωσης δίνονται στο σχήμα 4.3

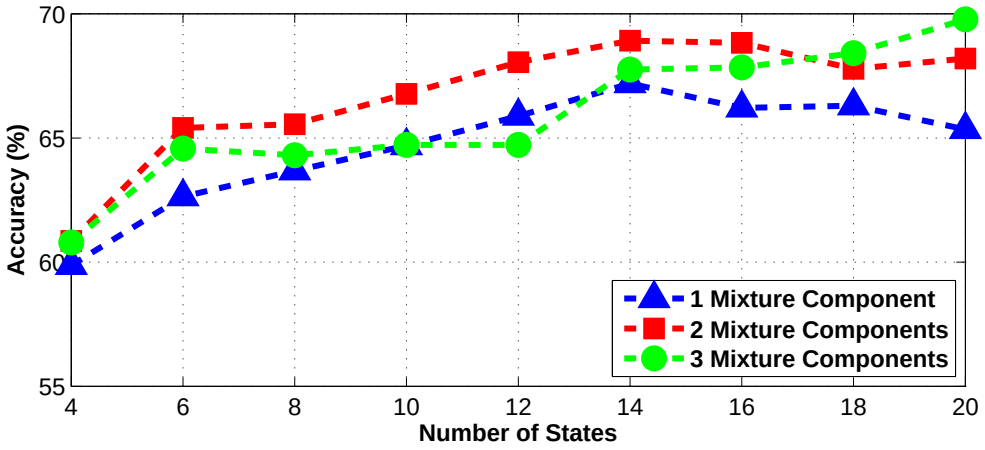


(α') Θέση Χεριού στον y-άξονα (Κατ. 1) (β') Απόσταση Χεριών στον y-άξονα (Κατ. 4)

Σχήμα 4.3: Το ιστόγραμμα αναπαριστά την κατανομή συγκεκριμένων χαρακτηριστικών Θέσης-Κίνησης σε διαφορετικές καταστάσεις των κρυφών Μαρκοβιανών μοντέλων. Η κόκκινη γραμμή, που έχει απεικονιστεί με υπέρθεση πάνω στο ιστόγραμμα, αντιστοιχεί στο μείγμα Γκαουσιανών (GMM) που έχει χρησιμοποιηθεί για να μοντελοποιήσει τη συγκεκριμένη κατανομή. Σε κάθε περίπτωση, η μοντελοποίηση είναι πολύ πιο ακριβής με τη χρήση δύο ή περισσότερων Γκαουσιανών.

Με δεδομένα τα όσα περιγράφηκαν παραπάνω, θα ερευνήσουμε ένα πλήθος από 4 έως 20 καταστάσεις ανά μοντέλο, ενώ για κάθε χαρακτηριστικό θα χρησιμοποιήσουμε από μία έως τρεις Γκαουσιανές για την περιγραφή της κατανομής ανά κατάσταση. Τα συγκεντρωτικά αποτελέσματα παρουσιάζονται στο σχήμα 4.4.

Με βάση αυτά τα αποτελέσματα είναι σαφής η υπεροχή της χρήσης περισσοτέρων από μία Γκαουσιανών για τη μοντελοποίηση κάθε κατάστασης.



Σχήμα 4.4: Ποσοστό επιτυχημένης ταξινόμησης για τα χαρακτηριστικά Θέσης-Κίνησης, με χρήση διαφορετικού αριθμού καταστάσεων, και διαφορετικού αριθμού Γκαουσιανών ανά κατάσταση, για τα κρυφά Μαρκοβιανά μοντέλα που εκπαιδεύουμε. Κάθε μία από τις τρεις διακεκομμένες γραμμές, αφορά χρήση μοντέλων με συγκεκριμένο αριθμό Γκαουσιανών ανά κατάσταση, ενώ η εξέλιξή τους στον x-άξονα παρουσιάζει τη διαφοροποίηση στην επίδοση για διαφορετικό αριθμό καταστάσεων. Όλα τα αποτελέσματα αφορούν τη βάση χειρονομιών ChaLearn.

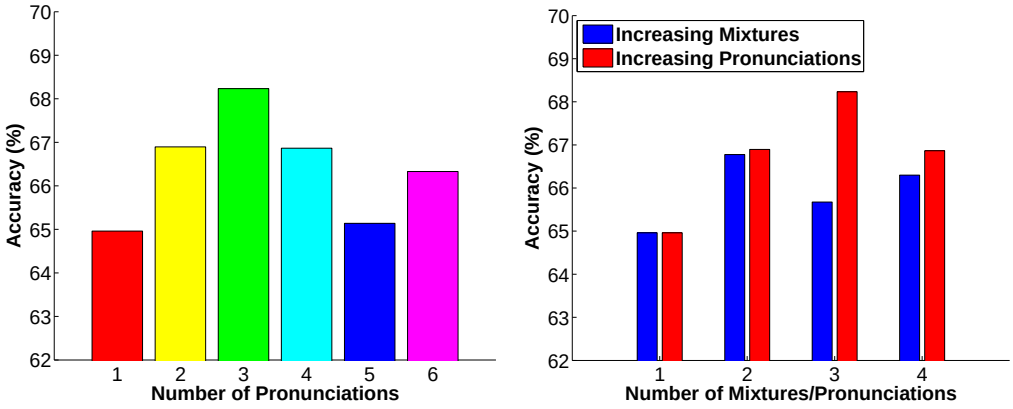
Μπορεί με χρήση δύο αντί τριών Γκαουσιανών τα αποτελέσματα να είναι εν γένει καλύτερα, αυτό όμως μπορεί να οφείλεται σε φαινόμενα υπερεκπαίδευσης (overfitting) στο σύνολο εκπαίδευσης, με αποτέλεσμα να μειώνεται η ικανότητα της γενίκευσης για τα μοντέλα μας. Όσον αφορά τώρα την επίδραση του αριθμού των καταστάσεων των μοντέλων μας, και πάλι, αντίστοιχα με τα χαρακτηριστικά της χειρομορφής, παρατηρείται εν γένει μια αύξηση των ποσοστών ταξινόμησης καθώς αυξάνονται οι καταστάσεις των μοντέλων. Υπάρχουν βεβαίως διαφοροποιήσεις σε κάθε περίπτωση (για χρήση τριών Γκαουσιανών, η τάση είναι συνεχώς ανοδική, ενώ αντίστοιχα για μία ή δύο Γκαουσιανές παρουσιάζεται στάσιμη κατάσταση για περισσότερες από 14 καταστάσεις), ωστόσο η συνολική εκτίμηση είναι ότι χρειάζονται τουλάχιστον 12 Γκαουσιανές για τη βέλτιστη μοντελοποίηση των χειρονομιών της βάσης μας. Συνολικά πάντως η βέλτιστη επίδοση σημειώνεται για 20 καταστάσεις, και τρεις Γκαουσιανές ανά κατάσταση, και ξεπερνάει το 69%.

Διερεύνηση "εκφορών" κάθε χειρονομίας

Όπως αναφέραμε και κατά την παρουσίαση της βάσης χειρονομιών ChaLearn, είναι πολύ συνηθισμένο να παρουσιάζονται ένα πλήθος διαφορετικών τρόπων εκτέλεσης της κάθε χειρονομίας. Για κάποια παραδείγματα, μπορεί να ανατρέξει κανείς στο σχήμα 1.1. Προκειμένου να αντιμετωπίσουμε ένα τέτοιο πρόβλημα,

θέλαμε έναν αυτόματο τρόπο να διαχωρίσουμε τις διαφορετικές "εκφορές" κάθε χειρονομίας, και στη συνέχεια να εκπαιδεύσουμε ένα διαφορετικό μοντέλο για κάθε εκφορά, ώστε να είναι όσο το δυνατόν καλύτερα εκπαιδευμένα τα μοντέλα μας (μιας και χρήση "ανάμεικτων" δεδομένων δημιουργεί προβλήματα στα μοντέλα). Για το σκοπό αυτό χρησιμοποιήσαμε τον αλγόριθμο Dynamic Time Warping (DTW), για matching των εκτελέσεων που ανήκουν στην ίδια "εκφορά".

Για τις ανάγκες του πειραματισμού μας επιλέγουμε κάθε φορά ένα συγκεκριμένο αριθμό "εκφορών" από δύο έως έξι για κάθε χειρονομία, και εκπαιδεύσαμε διαφορετικό μοντέλο για κάθε "εκφορά". Ο αριθμός των καταστάσεων των μοντέλων παρέμεινε σταθερός και ίσος με 13, ενώ χρησιμοποιήθηκε μία Γκαουσιανή ανά κατάσταση για όλα τα μοντέλα. Τα συγκεντρωτικά αποτελέσματα παρουσιάζονται στο σχήμα 4.5α΄.



(α΄) Αύξηση αριθμού "εκφορών" (β΄) Αριθμός "εκφορών" vs Γκαουσιανών

Σχήμα 4.5: Αριστερά: Ποσοστό επιτυχημένης ταξινόμησης για τα χαρακτηριστικά Θέσης-Κίνησης, καθώς αυξάνουμε το πλήθος των διαφορετικών "εκφορών" από μία έως έξι για κάθε χειρονομία. Δεξιά: Σύγκριση της επιτυχίας ταξινόμησης για αύξηση του αριθμού των "εκφορών" (μία Γκαουσιανή ανά κατάσταση για κάθε μοντέλο), σε σχέση με την αύξηση του αριθμού των Γκαουσιανών που περιγράφουν κάθε κατάσταση (μία εκφορά για κάθε λέξη-χειρονομία). Για αντικειμενικότητα σύγκρισης, όλα τα μοντέλα περιλαμβάνουν 13 καταστάσεις. Όλα τα αποτελέσματα αφορούν τη βάση χειρονομιών ChaLearn.

Αυτό που παρατηρούμε είναι ότι για χρήση μέχρι τριών "εκφορών" ανά χειρονομία έχουμε αύξηση των ποσοστών επιτυχημένης ταξινόμησης, ενώ από εκείνο το σημείο και μετά υπάρχει μικρή πτώση. Συνεπώς, όντως "κρύβονται" διαφορετικές εκφορές των χειρονομιών στο σύνολο δεδομένων μας (σχήμα 4.6), και είναι ενδιαφέρον ότι με έναν αυτόματο τρόπο (χρήση του αλγορίθμου DTW), καταφέραμε να διαχωρίσουμε τα δεδομένα μας σε διαφορετικές υποκλάσεις,

οδηγώντας τα επιμέρους μοντέλα "εκφορών" σε μεγαλύτερη επιτυχία ταξινόμησης.

Θα μπορούσε να επιχειρηματολογήσει κανείς βέβαια ότι η διάκριση "εκφορών", ίσως ισοδυναμεί με αύξηση του πλήθους των Γκαουσιανών στο μείγμα που αντιστοιχεί σε κάθε κατάσταση ενός μοντέλου. Κατά κάποιο τρόπο, η καλύτερη μοντελοποίηση κάθε κατάστασης (αντίστοιχη με αυτή που παρουσιάζεται στο σχήμα 4.3), ενδεχομένως θα μπορούσε να "συλλάβει" τις διαφοροποιήσεις στην εκτέλεση των χειρονομιών. Για μια κατάλληλη σύγκριση των δύο αντιμετωπίσεων, παραπέμπουμε στο σχήμα 4.5β΄. Εδώ έχουμε αντιπαραθέσει τα αποτελέσματα για προσθήκη μέχρι τεσσάρων "εκφορών" σε κάθε χειρονομία, με τα αντίστοιχα αποτελέσματα για την αύξηση των Γκαουσιανών στο μείγμα που μοντελοποιεί την κάθε κατάσταση. Όπως βλέπουμε, η χρήση των "εκφορών" οδηγεί ακόμα και σε καλύτερα αποτελέσματα, αν και οι επιδόσεις δεν απέχουν πολύ. Και όντως από τους γενικότερες πειραματισμούς μας εντοπίζουμε συγκρίσιμες επιδόσεις στις δύο περιπτώσεις (στην εκπαίδευση διαφορετικών μοντέλων για διαφορετικές εκφορές, και στη χρήση ενός ενιαίου μοντέλου, με την ενίσχυση του μείγματος των Γκαουσιανών που μοντελοποιούν κάθε κατάσταση). Δεν θα επιχειρηματολογήσουμε υπέρ της μίας ή της άλλης αντιμετώπισης, ωστόσο, χωρίς αμφιβολία, η αυτόματη διάκριση "εκφορών" με χρήση του αλγορίθμου DTW είναι μια ενδιαφέρουσα μεθοδολογία, που χρήζει εξέτασης σε παρόμοια προβλήματα.

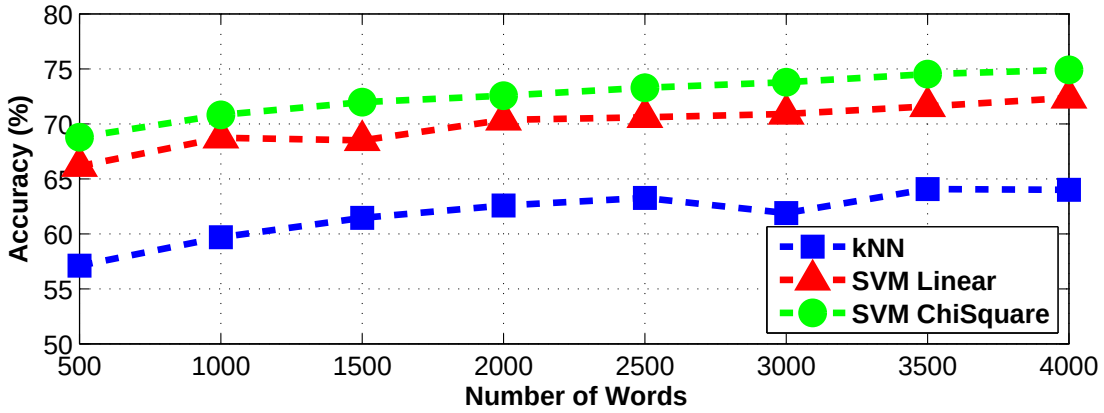
(α΄) (β΄) (γ΄) (δ΄)

Σχήμα 4.6: Πιθανές "εκφορές" που μπορούν να αναγνωριστούν αυτόματα με την εφαρμογή του αλγορίθμου DTW. (α΄) και (β΄): Χρήση αριστερού και δεξιού χεριού για εκτέλεση της χειρονομίας "vattene"· (γ΄) και (δ΄): Διαφοροποίηση της θέσης του χεριού (χαμηλά, ψηλά) για τη χειρονομία "vieniqui".

Χρήση ''στατικών'' ταξινομητών

Αντίστοιχα με την προσπάθειά μας στα χαρακτηριστικά της χειρομορφής, κάνουμε μία συμμετρική προσπάθεια ταξινόμησης και με τα χαρακτηριστικά θέσης-

κίνησης, με χρήση της τεχνικής Bag-of-Features, και "στατικών" ταξινομητών όπως kNN και SVM. Τα συνολικά αποτελέσματα, παρουσιάζονται στο σχήμα 4.7.



Σχήμα 4.7: Ποσοστά επιτυχημένης ταξινόμησης με τη χρήση της τεχνικής Bag-of-Features για χαρακτηριστικά θέσης-κίνησης και "στατικούς" ταξινομητές (kNN, SVM). Τα αποτελέσματα αφορούν τη βάση χειρονομιών ChaLearn.

Και πάλι, όπως και στα χαρακτηριστικά της χειρομορφής, το συγκεκριμένο πλαίσιο πειραματισμού, οδηγεί σε καλύτερα αποτελέσματα στο πρόβλημα της ταξινόμησης, με μέγιστο ποσοστό επιτυχημένης ταξινόμησης να πλησιάζει το 75% για χρήση SVM με χ^2-απόσταση. Αύξηση δηλαδή μεγαλύτερη από 5% σε σχέση με τα HMMs. Ωστόσο, αποφεύγουμε την ευρύτερη χρήση τέτοιων ταξινομητών, μιας και είναι σημαντική η αδυναμία που αναλύσαμε ήδη, και αφορά τη δυσκολία επέκτασης με άμεσο τρόπο σε προβλήματα αναγνώρισης.

Σύνοψη

Για μία σύνοψη των πειραματισμών στη βάση χειρονομιών ChaLearn, με χρήση χαρακτηριστικών Θέσης-Κίνησης, παραπέμπουμε στον πίνακα 4.1. Με μία σύντομη επισκόπηση, παρατηρούμε ότι επιτυγχάνουμε ποσοστά επιτυχίας για το πρόβλημα της ταξινόμησης, που πλησιάζουν το 70% με χρήση HMMs, ή πλησιάζουν το 75% με χρήση SVM. Και στις δύο περιπτώσεις, οι επιδόσεις είναι σαφώς βελτιωμένες σε σχέση με το κανάλι πληροφορίας της Χειρομορφής, σημειώνοντας κατά 10% υψηλότερα ποσοστά επιτυχημένης ταξινόμησης.

4.3.2 Πειράματα στη βάση Χειρονομιών MOBOT

Ένα μέρος από τους πειραματισμούς που έγιναν στη βάση χειρονομιών ChaLearn, μεταφέρθηκε και στη βάση χειρονομιών MOBOT. Για τούς λόγους που έχουμε ήδη

Data	**Feat.**	**Classifier**	**Exp. Type**	**Result**	**Comments**
ChaLearn	Various	HMM	Training-Validation	max 64.96%	13 states, 1 mixture.
ChaLearn	Θ-K	HMM	Training-Validation	max 69.78%	4-20 states, 1-3 mixtures.
ChaLearn	Θ-K	HMM	Training-Validation	max 68.23%	13 states, 1 mixture, 1-6 pronunciations.
ChaLearn	Θ-K+BoF	kNN/SVM	Training-Validation	max 74.93%	500-4000 visual words.

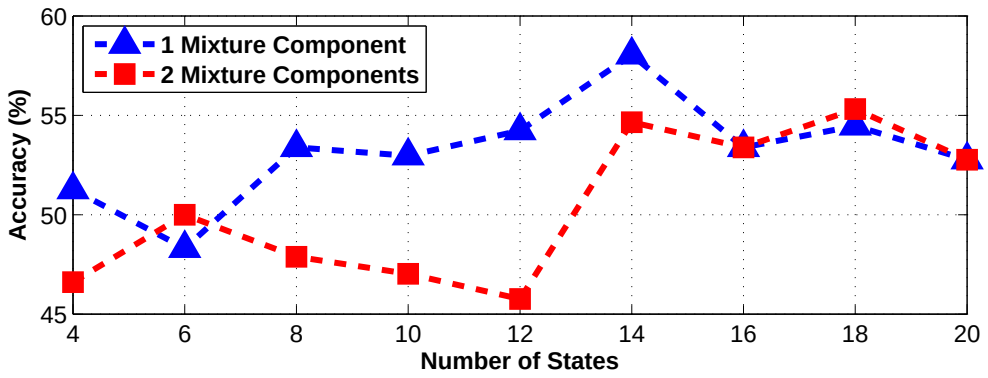
Πίνακας 4.1: Συνοπτική παρουσίαση των αποτελεσμάτων στη βάση χειρονομιών ChaLearn, χρησιμοποιώντας το κανάλι πληροφορίας της Θέσης-Κίνησης.

αναλύσει, περιοριστήκαμε σε χρήση και αξιολόγηση αντίστοιχων μεθόδων με αυτές που εργαστήκαμε και στη βάση ChaLearn. Και πάλι εργαστήκαμε με αντίστοιχο διάνυσμα χαρακτηριστικών (δουλεύοντας όμως μόνο με 2D χαρακτηριστικά, μιας και οι κινήσεις γίνονται κυρίως στο δισδιάστατο επίπεδο), ενώ ο πειραματισμός αφορούσε αποκλειστικά τη χρήση κρυφών Μαρκοβιανών μοντέλων. Για την αξιολόγηση εκτελέστηκαν πειράματα τύπου unseen signer, ώστε να εκτιμηθεί το μέγεθος της διακύμανσης στην εκτέλεση των χειρονομιών μεταξύ των διαφορετικών χρηστών.

Αξιολόγηση επίδρασης αριθμού καταστάσεων και πλήθους Γκαουσιανών ανά κατάσταση

Και πάλι μελετήσαμε την επίδραση του διαφορετικού αριθμού καταστάσεων στα μοντέλα μας, αλλά και το πλήθος των Γκαουσιανών κατανομών που χρησιμοποιούμε ανά κατάσταση. Εδώ είχαμε την ευκαιρία να χρησιμοποιήσουμε μέχρι δύο Γκαουσιανές ανά κατάσταση, λόγω του μικρού πλήθους δεδομένων που ήταν διαθέσιμο στη συγκεκριμένη βάση. Τα συγκεντρωτικά αποτελέσματα παρουσιάζονται στο σχήμα 4.8.

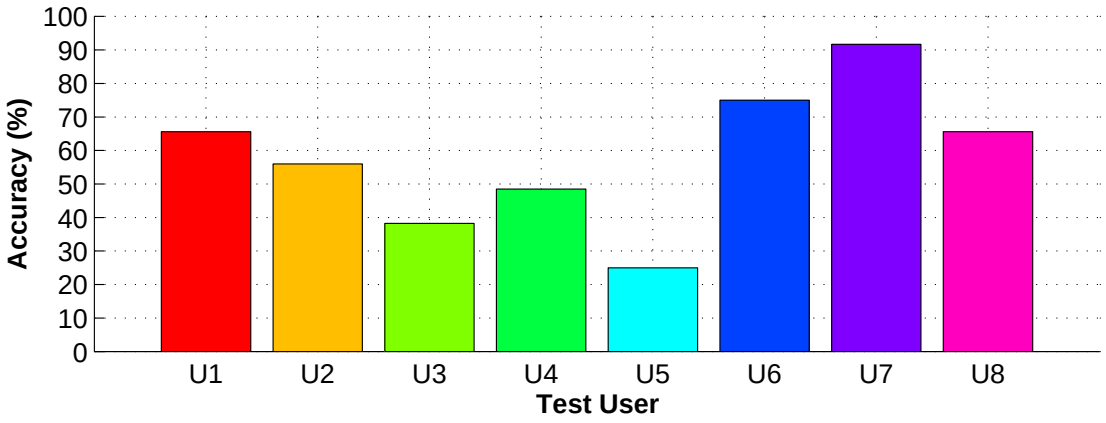
Ένα ενδιαφέρον συμπέρασμα προς σχολιασμό είναι ότι με την προσθήκη δεύτερης Γκαουσιανής κατανομής ανά κατάσταση, προκύπτει εν γένει μια πτώση στην επίδοση του συστήματος. Αυτό πιστεύουμε ότι οφείλεται στην υπερεκπαίδευση (overfitting), η οποία λόγω των λίγων διαθέσιμων παραδειγμάτων εκπαίδευσης είναι ακόμα πιο έντονη στη βάση MOBOT. Από την άλλη, όσον αφορά τον αριθμό καταστάσεων, η μέγιστη επίδοση φαίνεται να σημειώνεται για χρήση 14 καταστάσεων για κάθε μοντέλο, ξεπερνώντας το 58%, ενώ η περαιτέρω αύξηση ή μείωση του αριθμού καταστάσεων, οδηγεί σε χαμηλότερα ποσοστά επιτυχημένης αναγνώρισης.



Σχήμα 4.8: Ποσοστά επιτυχημένης ταξινόμησης, για διαφορετικό αριθμό καταστάσεων στα HMMs που χρησιμοποιούμε με χρήση της πληροφορίας Θέσης-Κίνησης. Τα αποτελέσματα αφορούν τη βάση MOBOT.

Αξιολόγηση επίδοσης ανά χρήστη

Εκτός από τα συγκεντρωτικά αποτελέσματα, έχει ενδιαφέρον να αξιολογήσουμε και την επίδοση σε κάθε χρήστη ξεχωριστά, με αντίστοιχο τρόπο, όπως κάναμε νωρίτερα και για τα χαρακτηριστικά της χειρομορφής. Σε αυτή την περίπτωση χρησιμοποιούμε 14 καταστάσεις ανά μοντέλο και μία Γκαουσιανή ανά κατάσταση (παραμετροποίηση που έδωσε τη μέγιστη επίδοση προηγουμένως), και απεικονίζουμε τα αποτελέσματα από τα unseen signer πειράματα στο σχήμα 4.9.



Σχήμα 4.9: Ποσοστά επιτυχημένης ταξινόμησης, για κάθε χρήστη ξεχωριστά. Σε κάθε περίπτωση, το σύστημα έχει εκπαιδευτεί με όλους τους χρήστες, πλην αυτού που χρησιμοποιείται για την αξιολόγηση (unseen signer πείραμα). Έχει γίνει χρήση των χαρακτηριστικών Θέσης-Κίνησης και HMM ταξινομητών, ενώ τα αποτελέσματα αφορούν τη βάση χειρονομιών MOBOT.

Αυτό που έχει ενδιαφέρον να παρατηρήσουμε στο συγκεκριμένο γράφημα, είναι η μεγάλη διακύμανση των ποσοστών επιτυχίας ανα χρήστη, με κάποιες περιπτώσεις η επίδοση να μην ξεπερνά το 30%, και με άλλες να υπερβαίνει το 90%. Αυτή τη μεγάλη διαφοροποίηση την αποδίδουμε στο γεγονός ότι υπάρχουν χρήστες που ενεργούν κατά μεγάλο ποσοστό ως outliers σε όλες τις χειρονομίες, τουλάχιστον σε ότι αφορά το κομμάτι της θέσης-κίνησης των χεριών. Ακόμα και αν η χειρομορφή είναι η σωστή/συνεπής, πραγματοποιούνται πολλά λάθη/ασυνέπειες στην κίνηση των χεριών των χρηστών (είτε λόγω διανοητικών, είτε λόγω σωματικών προβλημάτων). Συνεπώς, οι συγκεκριμένοι χρήστες κάνουν κατ' εξακολούθηση λάθη στην πλειοψηφία των χειρονομιών, με αποτέλεσμα να επηρεάζεται ιδιαίτερα η ταξινόμηση, όταν η εκπαίδευση έχει γίνει στους άλλους χρήστες.

Αυτές οι παρατηρήσεις, φαίνεται να δημιουργούν προβλήματα στους χρήστες 5 και 3 (και λιγότερο στον 4), και ευθύνονται κατά τη γνώμη μας για τα πολύ χαμηλά ποσοστά στις συγκεκριμένες περιπτώσεις. Αντίθετα οι υπόλοιποι χρήστες εμφανίζουν αρκετά υψηλά ποσοστά επιτυχημένης ταξινόμησης (συνήθως πάνω από 60%), που ειδικά για την περίπτωση του unseen signer πρόκειται για ιδιαίτερα αξιόλογη επίδοση.

Σύνοψη

Συνοπτικά, τα αποτελέσματά των πειραματισμών μας στη βάση χειρονομιών MOBOT, με χρήση των χαρακτηριστικών Θέσης-Κίνησης, παρουσιάζονται στον πίνακα 4.2. Σε αυτή την περίπτωση, σε αντίθεση με τα χαρακτηριστικά της Χειρομορφής (υποενότητα 3.4.3), οι επιδόσεις είναι σαφώς πιο ικανοποιητικές. Συγκεκριμένα, το υψηλότερο ποσοστό επιτυχημένης ταξινόμησης ξεπερνάει το 58% για τη μέθοδο αξιολόγησης unseen signer. Ασφαλώς, μας προβληματίζει η έντονη διακύμανση των ποσοστών επιτυχίας ανά χρήστη, ωστόσο μπορούμε να την αποδώσουμε στις ιδιαιτερότητες των χρηστών που έλαβαν μέρος στις λήψεις, οι οποίες θα απαιτούσαν και ειδικό χειρισμό των μεθόδων μας επί αυτών των δεδομένων.

Data	**Feat.**	**Classifier**	**Exp. Type**	**Result**	**Comments**
MOBOT	Θ-Κ	HMM	Unseen	max 58.05%	4-20 states, 1-2 mixtures.
MOBOT	Θ-Κ	HMM	Unseen	25% έως 91.67%	14 states, 1 mixture.

Πίνακας 4.2: Συνοπτική παρουσίαση των αποτελεσμάτων στη βάση χειρονομιών MOBOT, χρησιμοποιώντας το κανάλι πληροφορίας της Θέσης-Κίνησης.

4.4 Ανακεφαλαίωση αποτελεσμάτων χρήσης οπτικής πληροφορίας

Έχοντας ολοκληρώσει την παρουσίαση των μεθόδων εργασίας και των πειραματισμών με χρήση των οπτικών καναλιών πληροφορίας (κεφάλαια 3 και 4), και πριν εστιάσουμε στην παρουσίαση των σχημάτων σύμμειξης διαφορετικών ροών πληροφορίας (κεφάλαιο 5), επιχειρούμε μια ανακεφαλαίωση των πειραματισμών με χρήση αποκλειστικά ενός καναλιού πληροφορίας. Για το λόγο αυτό, δίνουμε συγκριτικά αποτελέσματα σε όλες τις διαθέσιμες βάσεις δεδομένων με χρήση είτε των χαρακτηριστικών Χειρομορφής, είτε των χαρακτηριστικών Θέσης-Κίνησης. Σε κάθε περίπτωση εστιάζουμε στα καλύτερα αποτελέσματα ανά πειραματισμό, και κάνουμε τις συγκρίσεις τόσο μεταξύ διαφορετικών τεχνικών στην ίδια βάση (για παράδειγμα διαφορετικό είδος χαρακτηριστικών ή διαφορετικό είδος ταξινομητών), αλλά και μεταξύ διαφορετικών βάσεων (για παράδειγμα επιδόσεις στη βάση ChaLearn, έναντι επιδόσεων στη βάση MOBOT).

Αρχικά, για τη βάση στατικών χειρομορφών, στην ουσία επαναλαμβάνουμε τα αποτελέσματα του πίνακα 3.1, μιας και δεν είναι δυνατό να χρησιμοποιηθούν χαρακτηριστικά Θέσης-Κίνησης λόγω της στατικής φύσης της βάσης. Παρ' όλα αυτά προχωράμε στην εκ νέου παρουσίαση των αποτελεσμάτων με χρήση χαρακτηριστικών Χειρομορφής στον πίνακα 4.3 για λόγους πληρότητας. Όπως και στη σύνοψη της υποενότητας 3.4.1, παρατηρούμε την πτώση των ποσοστών επιτυχημένης ταξινόμησης στο σχήμα αξιολόγησης unseen signer, σε σχέση με ένα τυπικό σχήμα χωρισμού 60%-40% του συνόλου των διαθέσιμων δεδομένων.

Exp. Type	HS
60%-40%	95.87%
Unseen	87.83%

Πίνακας 4.3: Συγκεντρωτικά αποτελέσματα στη βάση στατικών χειρομορφών, με χρήση οπτικών καναλιών πληροφορίας (αποκλειστικά κανάλι χειρομορφής, HS, για την εν λόγω βάση).

Στη συνέχεια, για τη βάση χειρονομιών ChaLearn συνδυάζουμε αποτελέσματα από τους πίνακες 3.2 και 4.1, και κάνουμε τη σύγκριση των επιδόσεων, τόσο σε σχέση με τα διαφορετικά κανάλια οπτικής πληροφορίας, όσο και σε σχέση με τους ταξινομητές διαφορετικών ειδών. Με τον πίνακα 4.4 συνεπώς, λαμβάνουμε μία καλύτερη, συγκεντρωτική εικόνα σε σχέση με τα ανεξάρτητα αποτελέσματα που είχαμε παρουσιάσει μέχρι τώρα. Όπως παρατηρούμε, τα χαρακτηριστικά της Θέσης-Κίνησης ενισχύουν κατά τουλάχιστον 10% την επιτυχημένη ταξινόμηση σε σχέση με τα χαρακτηριστικά Χειρομορφής. Επιπλέον, η χρήση SVM σε σχέση

με HMMs για το πρόβλημα της ταξινόμησης οδηγεί σε περίπου 5% υψηλότερα ποσοστά επιτυχίας.

Classifier	HS	MP
HMMs	59.58%	69.78%
SVM	64.13%	74.93%

Πίνακας 4.4: Συγκεντρωτικά αποτελέσματα επιτυχημένης ταξινόμησης στη βάση χειρονομιών ChaLearn, με χρήση οπτικών καναλιών πληροφορίας (χαρακτηριστικά χειρομορφής, HS, και χαρακτηριστικά θέσης-κίνησης, MP) και ταξινομητών διαφορετικού είδους.

Τέλος, για τη βάση χειρονομιών MOBOT, έχουμε συνδυάσει αποτελέσματα από τους πίνακες 3.3 και 4.2, με αποτέλεσμα να προκύψει ο πίνακας 4.5. Και εδώ παρατηρείται υπεροχή των χαρακτηριστικών Θέσης-Κίνησης έναντι αυτών της Χειρομορφής, και μάλιστα ακόμα μεγαλύτερη σε σχέση με ότι σημειώθηκε στη βάση ChaLearn (διαφορά υψηλότερη του 20%). Επίσης, συγκρίνοντας πάλι με τη βάση ChaLearn, τα ποσοστά επιτυχημένης ταξινόμησης είναι μικρότερα για τη βάση MOBOT, ωστόσο είναι σημαντικό να τονίσουμε ότι σε αυτή την περίπτωση έχει χρησιμοποιηθεί το σχήμα αξιολόγησης unseen signer, που είναι σαφώς πιο απαιτητικό από τα υπόλοιπα σχήματα που εξετάστηκαν.

Classifier	HS	MP
HMMs	37.71%	58.05%

Πίνακας 4.5: Συγκεντρωτικά αποτελέσματα επιτυχημένης ταξινόμησης στη βάση χειρονομιών MOBOT, με αποκλειστική χρήση οπτικών καναλιών πληροφορίας (χαρακτηριστικά χειρομορφής, HS, και χαρακτηριστικά θέσης-κίνησης, MP) και HMMs ταξινομητών.

Κεφάλαιο 5

Σύμμειξη ροών πληροφορίας

5.1 Γενικά

Μέχρι τώρα επικεντρωθήκαμε στην επεξεργασία αποκλειστικά μίας ροής πληροφορίας, είτε αυτή αφορούσε την εμφάνιση της χειρομορφής, είτε χαρακτηριστικά της θέσης και της κίνησης των χεριών κατά την εκτέλεση των χειρονομιών. Ωστόσο αποτελεί ακόμα μεγαλύτερη πρόκληση η σύμμειξη αυτών των διαφορετικών ροών πληροφορίας. Ειδικά στις συγκεκριμένες βάσεις δεδομένων που εργαζόμαστε οι οποίες είναι πολυτροπικές, και περιλαμβάνουν και την τροπικότητα του ήχου, πέρα από τις οπτικές ροές, το πρόβλημα της σύμμειξη διαφορετικών τροπικοτήτων γίνεται ένα αναγκαίο και σημαντικό κομμάτι της ερευνητικής δραστηριότητας.

Στο κεφάλαιο αυτό συνεπώς, παρουσιάζουμε την εργασία μας πάνω στη σύμμειξη των διαφορετικών διαθέσιμων τροπικοτήτων και καναλιών πληροφορίας, με χρήση της πολυτροπικής βάσης χειρονομιών ChaLearn. Πέρα από τα οπτικά κανάλια πληροφορίας, της χειρομορφής και την θέσης-κίνησης, σε αυτό το σημείο προστίθεται και η ηχητική πληροφορία. Διευκρινίζουμε ότι για τα πλαίσια αυτής της διπλωματικής που εστιάζει στην οπτική επεξεργασία, και στη σύμμειξη ροών πληροφορίας, δεν έχει διερευνηθεί πιο αναλυτικά το θέμα της επεξεργασίας της ηχητικής πληροφορίας και της αυτόματης αναγνώρισης ομιλίας. Αντίθετα, έχουν χρησιμοποιηθεί κλασικές μέθοδοι της βιβλιογραφίας για την επεξεργασία της τροπικότητας του ήχου, και την εκπαίδευση κατάλληλων μηχανών αναγνώρισης [93, 70].

Από την άλλη, σχετικά με τα κανάλια πληροφορίας της Χειρομορφής, και της Θέσης-Κίνησης, για την εξαγωγή χαρακτηριστικών ισχύουν τα όσα έχουμε αναφέρει στα αντίστοιχα κεφάλαια. Για την μοντελοποίηση, έχουμε κρατήσει σταθερά 13 καταστάσεις για τα μοντέλα και των δύο πληροφοριών, ενώ έχουμε χρησιμοποιήσει μία Γκαουσιανή ανά κατάσταση για την πληροφορία της Χειρομορφής, και πέντε

για την πληροφορία της Θέσης-Κίνησης.

Σε αντίθεση με τα πειράματα ταξινόμησης των ανεξάρτητων τροπικοτήτων, στο κεφάλαιο αυτό εστιάζουμε στο πρόβλημα της συνεχούς αναγνώρισης χειρονομιών, οπότε και φαίνεται η δύναμη των HMMs, τα οποία χρησιμοποιούνται κατ' αποκλειστικότητα. Για τα πειράματα αναγνώρισης χρησιμοποιούμε για την εκπαίδευση τόσο το Training Set, όσο και το Validation Set της ChaLearn βάσης, ενώ απομένει μόνο το Evaluation Set για τους σκοπούς της αξιολόγησης.

Τα κύρια σχήματα σύμμειξης που θα μας απασχολήσουν στη συνέχεια είναι η τεχνική του N-Best List Rescoring (στην οποία θα αναφερόμαστε και ως $P1$), τα PaHMMs (ή $P2$), καθώς και ένας συνδυασμός αυτών των επιμέρους σχημάτων. Θα κάνουμε μία σύντομη παρουσίαση του κάθε σχήματος, και στο τέλος θα προχωρήσουμε στην παράθεση αποτελεσμάτων τόσο με χρήση των τροπικοτήτων ανεξάρτητα, όσο και με εκμετάλλευση των σχημάτων σύμμειξης, ώστε να αξιολογήσουμε τη συνεισφορά τους στο συνολικό σύστημα αναγνώρισης.

5.2 N-Best List Rescoring ($P1$)

Η αξιολόγηση των N-καλύτερων υποθέσεων είναι μια τεχνική που προέρχεται από την κοινότητα της επεξεργασίας ομιλίας και φυσικού λόγου. Χρησιμοποιήθηκε για πρώτη φορά για το συνδυασμό της ομιλίας με τη φυσική γλώσσα από τους Chow et al. [15] και στη συνέχεια χρησιμοποιήθηκε ξανά από τους Ostendorf et al. για την ενσωμάτωση διαφορετικών τεχνικών αναγνώρισης [61].

Στην περίπτωσή μας το N-Best List Rescoring θεωρήθηκε ως μια αποδοτική μέθοδος για το συνδυασμό των διαφορετικών ροών πληροφορίας. Εν συντομία, η ιδέα πίσω από αυτή τη μέθοδο έγκειται στο ότι ο αλγόριθμος Viterbi δεν επιστρέφει την καλύτερη ακολουθία αναγνώρισης, αλλά μία λίστα από τις N-καλύτερες ακολουθίες (υποθέσεις). Συνεπώς έχουμε τη δυνατότητα να επαναξιολογήσουμε αυτή τη λίστα και με τις υπόλοιπες τροπικότητες, μια αξιολόγηση, που συνήθως οδηγεί σε αναδιάταξή της. Με την πεποίθηση ότι η λίστα μετά την αναδιάταξη είναι πιο "αξιόπιστη" σε σχέση με την αρχική, εφόσον όλες οι τροπικότητες έχουν συμμετέχει στην κατασκευή της, βασιζόμαστε σε αυτή για την επιλογή του αποτελέσματος της αναγνώρισης.

Έχοντας παρουσιάσει το κίνητρο πίσω από τη χρήση του N-Best List Rescoring, αλλά και μετά από μια πιο υψηλού επιπέδου περιγραφή, εστιάζουμε στη συνέχεια στην πιο αναλυτική περιγραφή της μεθόδου που εφαρμόζουμε.

Μέθοδος

Αρχικά, με χρήση του αλγορίθμου Viterbi, παράγουμε μία λίστα από τις N-καλύτερες υποθέσεις $H_1, \ldots, H_N$, κάθε μία από τις οποίες περιλαμβάνει μια

ακολουθία από χειρονομίες-λέξεις που αναγνωρίστηκαν. Εφόσον κάθε υπόθεση αποτελείται από μια ακολουθία χειρονομιών, θεωρούμε ότι είναι της μορφής $H_i = [g_1 g_2 \dots g_M]$, όπου με g_i συμβολίζουμε τη χειρονομία που αναγνωρίστηκε στην i θέση. Παράλληλα, από τον αλγόριθμο Viterbi έχει προκύψει και το Viterbi σκορ για κάθε μία από αυτές τις υποθέσεις:

$$v_i^m = \max_{q \in Q} \log P(O_m, q | H_i, \lambda), i = 1, \dots, N, \tag{5.1}$$

όπου O_m είναι η παρατηρούμενη ακολουθία για κάθε τροπικότητα m, q είναι η ακολουθία καταστάσεων όλων των πιθανών ακολουθιών στο Q, και λ είναι το αντίστοιχο σύνολο μοντέλων.

Έχοντας διαθέσιμες τις παραπάνω υποθέσεις, τις επαναξιολογούμε και για τις υπόλοιπες ροές πληροφορίας, ώστε να προκύψει το Viterbi σκορ για κάθε μία από τις διαθέσιμες ροές. Στην ουσία, επιβάλλουμε και στις υπόλοιπες τροπικότητες να αναγνωρίσουν κάθε μία από τις υποθέσεις H_i (διαδικασία (force-alignment), και λαμβάνουμε το σκορ που υπολογίζεται από τον αλγόριθμο Viterbi για τη συγκεκριμένη ακολουθία αναγνώρισης. Ο συνδυασμός αυτών των σκορ, αναδιατάσσει την αρχική λίστα των H_i υποθέσεων, και από εκεί προκύπτει η ανανεωμένη καλύτερη υπόθεση η οποία επιλέγεται ως η επικρατέστερη για την αναγνώριση της συγκεκριμένης ακολουθίας.

Στα πλαίσια της συγκεκριμένης εργασίας μας, η αρχική λίστα υποθέσεων προκύπτει από την τροπικότητα του ήχου, η οποία φαίνεται να σημειώνει τα καλύτερα αποτελέσματα σε επιμέρους πειράματα που έγιναν για κάθε ροή πληροφορίας ανεξάρτητα (πίνακας 5.1) και θεωρείται ως η πιο αξιόπιστη. Στη συνέχεια, η λίστα υποθέσεων επαναξιολογείται από τη ροή πληροφορίας της χειρομορφής, και από αυτή της θέσης-κίνησης και ακολουθεί ένας γραμμικός συνδυασμός των επιμέρους σκορ Viterbi, σύμφωνα με την επόμενη σχέση:

$$v_i^{p1} = \sum_m w_m^{p1} v_i^m \tag{5.2}$$

όπου v_i^m είναι το σκορ Viterbi για την υπόθεση H_i με βάση την τροπικότητα m και w_m^{p1} είναι το αντίστοιχο βάρος για την ίδια τροπικότητα. Τα προαναφερθέντα βάρη w_m^{p1} έχουν επιλεγεί με σκοπό να βελτιστοποιούν την αναγνώριση, και η στάθμισή τους έχει γίνει με χρήση του validation set. Τελικά, η υπόθεση με το μεγαλύτερο συνδυασμένο σκορ, και η αντίστοιχη ακολουθία λέξεων-χειρονομιών είναι η πλέον πιθανή μετά την ολοκλήρωση αυτού του βήματος, και αυτή που αναγνωρίζεται για τη δεδομένη ακολουθία.

5.3 Parallel HMMs ($P2$)

Τα παράλληλα HMMs (PaHMMs), προτάθηκαν από τους Vogler και Metaxas για το πρόβλημα της αναγνώρισης νοηματικής γλώσσας [85], με σκοπό τον συνδυασμό διαφορετικών καναλιών πληροφορίας. Τα PaHMMs μοντελοποιούν C κανάλια πληροφορίας, χρησιμοποιώντας και C ανεξάρτητα HMMs με διαφορετική έξοδο για το καθένα από αυτά. Έτσι, εφόσον δεν υπάρχει κάποια επιρροή κάποιας ροής πληροφορίας πάνω στις υπόλοιπες, τα PaHMMs μπορούν να θεωρηθούν ως τυπικά HMMs, τα οποία όμως χρησιμοποιούνται παράλληλα.

Στο πλαίσιο των προβλημάτων που εξετάζουμε, τα PaHMMs χρησιμοποιούνται για την λήψη απόφασης σε προβλήματα ταξινόμησης, συνεπώς θα χρειαστεί να κάνουμε κάποιο επιπλέον βήμα για να τα χρησιμοποιήσουμε στο πρόβλημα της αναγνώρισης σε συνεχές οπτικοακουστικό υλικό. Η τακτική που εφαρμόσαμε σε αυτή την περίπτωση ήταν η χρονική κατάτμηση του συνεχούς βίντεο σε επιμέρους αποσπάσματα, και εν συνεχεία η ταξινόμηση τους σε μία από τις διαθέσιμες κλάσεις με αξιοποίηση των PaHMMs. Η πιο φυσιολογική απόφαση για την αυτόματη χρονική κατάτμηση του πολυτροπικού σήματος, ήταν με βάση την πληροφορία για τα χρονικά όρια των λέξεων που προκύπτει από τον αλγόριθμο Vitebi. Επιλέγοντας ως επί το πλείστον το αποτέλεσμα της εφαρμογής του αλγορίθμου Viterbi στην ηχητική πληροφορία (θεωρώντας την ως την πλέον αξιόπιστη), μπορούμε να λάβουμε μία χρονική κατάτμηση του πολυτροπικού σήματος. Εν συνεχεία, αυτά τα ανεξάρτητα οπτικοακουστικά τμήματα μπορούν να ταξινομηθούν στην πιο πιθανή κατηγορία με χρήση των PaHMMs.

Και πάλι, έχοντας παρουσιάσει μία υψηλού επιπέδου εικόνα για την εφαρμογή των PaHMMs στο σύστημά μας, συνεχίζουμε με την αλγοριθμική περιγραφή της μεθόδου που ακολουθείται.

Μέθοδος

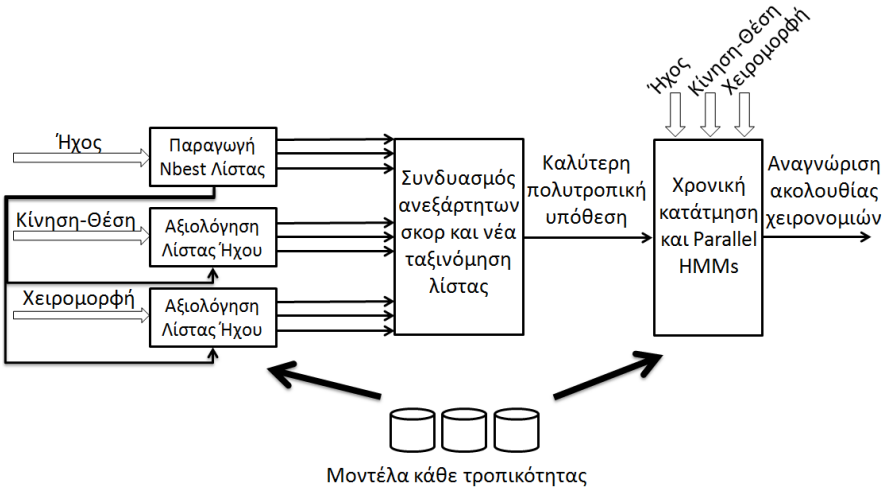
Στόχος των PaHMMs είναι η ταξινόμηση ενός οπτικοακουστικού αποσπάσματος s λαμβάνοντας όμως υπόψη όλα τα διαφορετικά κανάλια πληροφορίας. Με αυτό το κίνητρο, για το συγκεκριμένο απόσπασμα s και για κάθε κανάλι πληροφορίας m υπολογίζουμε τη λογαριθμική πιθανότητα $LL^m_{s,j} = \max_{q \in Q} \log P(O_m, q|\lambda^m_j)$, όπου λ^m_j είναι οι παράμετροι του HMM για τη λέξη-χειρονομία j το οποίο έχει εκπαιδευτεί με δεδομένα της ροής πληροφορίας m, ενώ q είναι η ακολουθία καταστάσεων. Τέλος, προκύπτει ένας γραμμικός συνδυασμός των $LL^m_{s,j}$ για όλα τα διαφορετικά κανάλια πληροφορίας, που οδηγεί στο τελικό σταθμισμένο σκορ:

$$LL^{p2}_{s,j} = \sum_m w^{p2}_m LL^m_{s,j}, \tag{5.3}$$

όπου w_m^{p2} είναι το βάρος για το κανάλι πληροφορίας, και το οποίο έχει εκπαιδευτεί στο validation set. Η χειρονομία-λέξη που έχει το υψηλότερο σκορ μετά τον παραπάνω συνδυασμό, είναι και αυτή που αναγνωρίζεται στο συγκεκριμένο απόσπασμα s.

Συνδυασμός με N-Best List Rescoring

Μία επέκταση στα όσα αναφέρθηκαν παραπάνω, είναι ο συνδυασμός των δύο σχημάτων σύμμειξης, του N-Best List Rescoring, και των PaHMMs. Σε αυτό το συνδυασμό, αντί της αρχικής χρονικής κατάτμησης που χρησιμοποιήσαμε, και η οποία προέρχεται από την πιο πιθανή υπόθεση της τροπικότητας του ήχου (απλό $P2$), επιλέγουμε να χρησιμοποιήσουμε τη χρονική κατάτμηση που επιστρέφει η πληροφορία του ήχου, με βάση όμως την πλέον πιθανή υπόθεση που προέκυψε από την έξοδο του σχήματος $P1$ ($P1 + P2$). Έτσι, η χρονική κατάτμηση προέρχεται και πάλι από την τροπικότητα του ήχου, αλλά όλες οι ροές πληροφορίας έχουν συμβάλλει στην απόφαση για το ποια υπόθεση θα χρησιμοποιηθεί (με forced alignment) για την χρονική κατάτμηση του πολυτροπικού σήματος. Προς επισκόπηση, ένα συνολικό block διάγραμμα του συνδυασμού των μεθόδων σύμμειξης $P1 + P2$ παρουσιάζεται στο σχήμα 5.1.



Σχήμα 5.1: Block Diagram που απεικονίζει το συνδυασμό των μεθόδων σύμμειξης, $P1 + P2$. Τα οπτικά κανάλια πληροφορίας αξιοποιούνται στο πρώτο βήμα για την επαναξιολόγηση της λίστας των N-καλύτερων υποθέσεων που προκύπτουν από την τροπικότητα του ήχου. Η καλύτερη υπόθεση που προκύπτει από την ανανεωμένη λίστα τροφοδοτεί το δεύτερο στάδιο, των PaHMMs, μετά από κατάτμηση του πολυτροπικού σήματος. Το αποτέλεσμα του δεύτερου σταδίου συνθέτει την ακολουθία χειρονομιών-λέξεων που αναγνώρισε το σύστημά μας.

5.4 Πειραματικά αποτελέσματα

Χρησιμοποιώντας τα σχήματα σύμμειξης που περιγράφηκαν νωρίτερα, προχωρήσαμε σε εφαρμογή τους στην πολυτροπική βάση χειρονομιών ChaLearn, με σκοπό την αξιολόγησή τους. Στις επόμενες υποενότητες παρουσιάζουμε ένα μέρος του πειραματισμού που έγινε, με σκοπό την αξιολόγηση των σχημάτων σύμμειξης, τόσο μεταξύ τους, όσο και σε σχέση με τα αποτελέσματα των επιμέρους ανεξάρτητων ροών πληροφορίας.

5.4.1 Αναγνώριση με βάση μία ροή πληροφορίας

Αρχικά παρουσιάζουμε τα αποτελέσματα της αναγνώρισης, χρησιμοποιώντας την κάθε ροή πληροφορίας ανεξάρτητα από τις υπόλοιπες. Αυτά τα αποτελέσματα αποτελούν μία αρχική βάση που θα πρέπει να βελτιωθεί με το συνδυασμό των ροών πληροφορίας. Το ποσοστό επιτυχημένης αναγνώρισης για κάθε μία από τις τρεις ροές πληροφορίας (ήχος, θέση-κίνηση και χειρομορφή), παρουσιάζεται στον πίνακα 5.1.

Αυτό που παρατηρούμε εδώ είναι η μεγάλη υπεροχή της πληροφορίας του ήχου, η οποία και γι' αυτό άλλωστε χρησιμοποιείται ως βάση στα περισσότερα σχήματα σύμμειξης που χρησιμοποιούμε. Από την άλλη η πληροφορία της χειρομορφής είναι αυτή που εμφανίζει τα χαμηλότερα ποσοστά επιτυχίας. Ήταν αναμενόμενο ότι θα είχε χαμηλότερα ποσοστά αναγνώρισης, όπως είχε και χαμηλότερα ποσοστά ταξινόμησης σε σχέση με την πληροφορία Θέσης-Κίνησης, όμως η διαφορά σε πειράματα αναγνώρισης είναι αρκετά μεγαλύτερη. Την "αδυναμία" αυτή την αποδίδουμε στο γεγονός ότι τα χαρακτηριστικά της χειρομορφής δεν "συλλαμβάνουν" με ξεκάθαρο τρόπο πληροφορία για κίνηση ή ακινησία, οπότε δεν είναι εύκολο να διαχωριστούν αυτές οι καταστάσεις. Με αυτό το σκεπτικό, η αναγνώριση, που περιλαμβάνει και εντοπισμό της χειρονομίας εκτός από ταξινόμησή της, είναι ακόμα πιο δύσκολο να λειτουργήσει με επιτυχία. Παρ' όλα αυτά γνωρίζοντας την ικανοποιητική επίδοση των χαρακτηριστικών της χειρομορφής σε προβλήματα ταξινόμησης, έχουμε την ένδειξη ότι θα λειτουργήσουν ικανοποιητικά στο σχήμα των PaHMMs.

5.4.2 Αναγνώριση με χρήση των σχημάτων σύμμειξης

Το επόμενο βήμα, που αποτελεί και τον κύριο στόχο αυτό του κεφαλαίου, είναι να πειραματιστούμε με τη χρήση των σχημάτων σύμμειξης, και να αξιολογήσουμε την επίδοσή τους. Με βάση την ανάλυση που κάναμε μέχρι τώρα, ξεχωρίζουν τρεις περιπτώσεις που έχουν ιδιαίτερο ενδιαφέρον. Η εφαρμογή του σχήματος $P1$, η εφαρμογή του σχήματος $P2$, και η συνδυασμένη εκδοχή των δύο σχημάτων, $P1 + P2$. Τα αποτελέσματα για αυτές τις περιπτώσεις, παρουσιάζονται στον πίνακα 5.1.

Single Modalities			**Fusion**		
Aud.	*MP*	*HS*	$P1$	$P2$	$P1+P2$
78.4	47.6	13.3	85.8	87.2	88.2

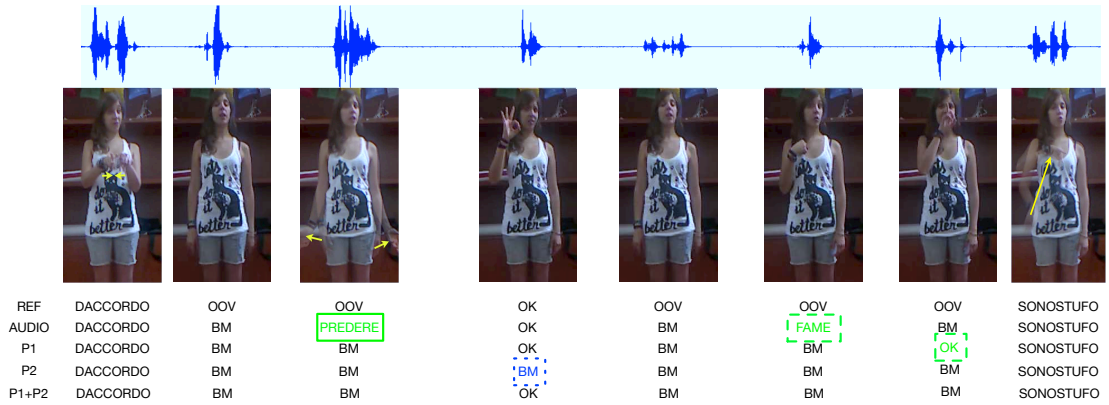
Πίνακας 5.1: Αξιολόγηση της επίδοσης των ανεξάρτητων τροπικοτήτων (Single Modalities, συμπεριλαμβανομένων του ήχου (Aud.), της θέσης-κίνησης (MP), και της χειρομορφής (HS), καθώς και των διαφορετικών σχημάτων σύμμειξης που προτείνουμε.

Όπως παρατηρούμε με τα σχήματα σύμμειξης που χρησιμοποιήσαμε πετύχαμε αύξηση από 7.4% μέχρι 9.8%, σε σχέση με την καλύτερη επίδοση της αναγνώρισης με χρήση ανεξάρτητων τροπικοτήτων. Ιδιαίτερα, παρατηρούμε ότι το $P2$ έχει λίγο βελτιωμένη επίδοση σε σχέση με το $P1$, ενώ ο συνδυασμός τους, $P1+P2$, πετυχαίνει την καλύτερη επίδοση για το συγκεκριμένο πρόβλημα.

Παράδειγμα Αναγνώρισης

Το κίνητρο για τη συνδυασμένη χρήση των δύο σχημάτων σύμμειξης, επιβεβαιώνεται πέρα από τα ποσοστά αναγνώρισης, και από παραδείγματα της εκτέλεσης της μεθόδου σε διάφορες ακολουθίες χειρονομιών της βάσης μας. Ένα τέτοιο παράδειγμα παρουσιάζουμε προς επισκόπηση στο σχήμα 5.2.

Αναφερόμενοι σε αυτό το παράδειγμα, βλέπουμε ότι η τροπικότητα του ήχου έχει εισαγάγει λανθασμένα δύο λέξεις-χειρονομίες, το PREDERE και το FAME. Ωστόσο, μετά την εφαρμογή των σχημάτων σύμμειξης τα λάθη αυτά εξαλείφονται, αφού η οπτική πληροφορία απορρίπτει την εμφάνιση αυτών των λέξεων στον τελικό συνδυασμό. Αυτό είναι ένα από τα επιθυμητά αποτελέσματα των σχημάτων σύμμειξης· αν σε μία τροπικότητα υπάρχει αμφιβολία σε κάποια περίπτωση, αλλά σε έταιρη τροπικότητα, η απόφαση που λαμβάνεται είναι ξεκάθαρη, τότε πιθανότατα θα υπερισχύσει με χρήση ενός αποδοτικού σχήματος σύμμειξης. Μπορεί δηλαδή τα δύο σχήματα σύμμειξης να εξάλειψαν τα δύο αυτά λάθη, ωστόσο το καθένα παίρνει μία λανθασμένη απόφαση, το $P1$ εισάγει τη χειρονομία OK σε λανθασμένη θέση, ενώ το $P2$ προχωράει σε διαγραφή της χειρονομίας OK. Αυτά τα λάθη, εξαλείφονται οριστικά στο συνδυασμό $P1+P2$, όπου όπως βλέπουμε αναγνωρίζει χωρίς κάποιο λάθος τη συγκεκριμένη ακολουθία. Έτσι, δικαιολογείται η συνδυασμένη χρήση των δύο σχημάτων, εφόσον με αυτό τον τρόπο καταφέραμε να εξαλείψουμε λάθη, τα οποία εισήγαγαν τα δύο σχήματα ανεξάρτητα (αλλά και οι ανεξάρτητες τροπικότητες).



Σχήμα 5.2: Παράδειγμα αναγνώρισης για μία ακολουθία λέξεων-χειρονομιών. Στην κορυφή τοποθετείται η οπτικοποίηση του σήματος ήχου για το συγκεκριμένο απόσπασμα και στη συνέχεια η οπτική πληροφορία, μέσω μιας σειράς στιγμιοτύπων του καναλιού RGB. Ακολουθεί η πραγματική ακολουθία λέξεων-χειρονομιών για το παράδειγμα (REF), τα αποτελέσματα αναγνώρισης για την τροπικότητα του ήχου (AUDIO), και για τα τρία σχήματα σύμμειξης (P1, P2, P1+P2). Το background μοντέλο bm μοντελοποιεί τις λέξεις εκτός λεξιλογίου (OOV).

Συμμετοχή στο Διαγωνισμό Αναγνώρισης Χειρονομιών

Η μεγάλη επιτυχία της συγκεκριμένης προσέγγισης, εξακριβώνεται και από την επίδοση που σημειώνει σε σχέση με τις υπόλοιπες μεθόδους που προτάθηκαν στα πλαίσια του πρόσφατου διαγωνισμού πολυτροπικής αναγνώρισης χειρονομιών. Τα πλήρη αποτελέσματα του συγκεκριμένου διαγωνισμού, σε σύγκριση με τη δική μας μέθοδο παρουσιάζονται στον πίνακα 5.2.

Όπως βλέπουμε, η μέθοδος που προτείνουμε είναι σαφώς βελτιωμένη, τόσο σε σχέση με τις μεθόδους των ομάδων που σημειώνουν τις υψηλότερες επιδόσεις, όσο και σε σχέση με την αρχική μέθοδο που είχαμε υποβάλλει στον ίδιο διαγωνισμό, και η οποία είχε καταλάβει την 5η θέση [28]. Αξίζει να σημειωθεί ότι η βελτίωση στην επίδοση της μεθόδου μας, οφείλεται αφενός στη χρήση του validation set για την εκπαίδευση του συστήματός μας (σύνολο το οποίο από εμάς είχε χρησιμοποιηθεί αποκλειστικά για τη ρύθμιση των παραμέτρων στην αρχική προσπάθεια, σε αντίθεση με τις υπόλοιπες ομάδες), και αφετέρου στην προσθήκη του σχήματος $P2$, το οποίο δεν είχε ενσωματωθεί στο αρχικό σύστημα.

Συμπέρασμα

Το συμπέρασμα που προκύπτει από τα παραπάνω αποτελέσματα, είναι προφανώς η υπεροχή που έχει η σύμμειξη ροών πληροφορίας σε σχέση με την επεξεργασία ανεξάρτητα της κάθε μίας από αυτές. Και μπορεί αυτό να ήταν κάτι αναμενόμενο,

Rank	Approach	Lev. Dist.	Acc.%	RER
-	Our [64]	0.11802	88.198	-
1	iva.mm [95]	0.12756	87.244	+7.48
2	wweight	0.15387	84.613	+23.30
3	E.T. [5]	0.17105	82.895	+31.00
4	MmM	0.17215	82.785	+31.44
5	pptk	0.17325	82.675	+31.88

Πίνακας 5.2: Η προσέγγισή μας σε σχέση με τις μεθόδους που κατέλαβαν τις πέντε πρώτες θέσεις στον αντίστοιχο διαγωνισμό αναγνώρισης χειρονομιών. Συμπεριλαμβάνεται η ακρίβεια της αναγνώρισης (Acc.%), η απόσταση Levenshtein (Lev. Dist., το μετρικό που χρησιμοποιήθηκε στο διαγωνισμό και ισούται με $1 - Acc.$), καθώς και τη σχετική μείωση λάθους.

ωστόσο δεν θα πρέπει να παραβλέπουμε και τη μεγάλη σημασία που καταλαμβάνει ο τρόπος που θα γίνει η εν λόγω σύμμειξη. Υπάρχει πληθώρα μεθόδων που χρησιμοποιούνται στη βιβλιογραφία, αλλά η επιλογή των κατάλληλων, και ο τρόπος που θα εφαρμοστούν ή συνδυαστούν παραμένει ένα δύσκολο πρόβλημα. Στην περίπτωσή μας, κρίνοντας τόσο από το πολύ υψηλό ποσοστό επιτυχίας, όσο και από το γεγονός ότι η μέθοδός μας ξεπερνάει το state-of-the-art που επιτεύχθηκε στο διαγωνισμό με τις ίδιες ακριβώς συνθήκες πειραματισμού, συμπεραίνουμε ότι εφαρμόσαμε ένα ιδιαίτερα επιτυχημένο συνδυασμό σχημάτων σύμμειξης.

Κεφάλαιο 6

Σύνοψη

Το κεφάλαιο αυτό αποτελεί την κατακλείδα της παρούσας διπλωματικής, και αξιοποιείται ώστε αφενός να παρουσιάσει μια ανακεφαλαίωση των βασικών συνεισφορών της διπλωματικής αυτής, και αφετέρου προκειμένου να διατυπώσει κάποιες ιδέες που θα μπορούσαν να αποτελέσουν κίνητρο για σχετική έρευνα, πέρα από τα όρια αυτής της εργασίας.

6.1 Ανακεφαλαίωση-Συνεισφορά

Στο σημείο αυτό κάνουμε μια σύνοψη της έρευνας που πραγματοποιήθηκε στα πλαίσια της διπλωματικής στο πεδίο της αναγνώρισης χειρονομιών. Εστιάζουμε στα βασικά σημεία, όπως αυτά προέκυψαν από την παρουσίαση των προηγούμενων κεφαλαίων. Οι κύριες πτυχές της έρευνας της παρούσας διπλωματικής, συνοψίζονται ως εξής:

- Στην πορεία της διπλωματικής χρησιμοποιήσαμε τρεις διαφορετικές βάσεις δεδομένων για την έρευνα και τον πειραματισμό μας, τη βάση στατικών χειρομορφών που οι ίδιοι καταγράψαμε, την πολυτροπική βάση χειρονομιών ChaLearn, και την επίσης πολυτροπική αλλά και πολυ-αισθητηριακή βάση χειρονομιών MOBOT.

- Χρησιμοποιήσαμε ισχυρές μεθόδους της αναγνώρισης προτύπων για την μοντελοποίηση και την αναγνώριση χειρονομιών, όπως τα κρυφά Μαρκοβιανά μοντέλα, αλλά και άλλοι δημοφιλείς ταξινομητές, όπως τα SVM και kNN.

- Εξετάσαμε αναλυτικά τη ροή πληροφορίας της χειρομορφής, με την εξαγωγή πλήθους διαφορετικών περιγραφητών σε αυτή, αλλά και μέσω εκτενούς πειραματισμού.

- Εστιάσαμε επιπλέον στη ροή πληροφορίας της θέσης-κίνησης, χρησιμοποιώντας χαρακτηριστικά από τη θέση και την κίνηση των χεριών, για την μοντελοποίηση των χειρονομιών. Αντίστοιχα με την πληροφορία της χειρομορφής, προχωρήσαμε σε εκτενή πειραματισμό, τόσο όσον αφορά το είδος των χαρακτηριστικών, όσο και σχετικά με τη μοντελοποίηση που θα χρησιμοποιήσουμε (πχ χρήση διαφορετικών "εκφορών").

- Εξετάσαμε το πρόβλημα της πολυτροπικής αναγνώρισης χειρονομιών, μέσω δύο διαφορετικών σχημάτων σύμμειξης, αλλά και συνδυασμό αυτών. Προχωρήσαμε σε ανάλυση των αποτελεσμάτων, τόσο με χρήση των ανεξάρτητων τροπικοτήτων, όσο και με χρήση των διαφορετικών σχημάτων.

- Επιτύχαμε ποσοστό αναγνώρισης στη βάση δεδομένων ChaLearn που ξεπερνάει τις επιδόσεις των αντίστοιχων προσεγγίσεων στα πλαίσια του πρόσφατου διαγωνισμού πολυτροπικής αναγνώρισης χειρονομιών.

6.2 Μελλοντικές Κατευθύνσεις

Παρότι η συγκεκριμένη διπλωματική εξέτασε αναλυτικά το πρόβλημα της πολυτροπικής αναγνώρισης χειρονομιών, με βάση τα παραπάνω συμπεράσματα, αλλά και το όλο υλικό που παρατίθεται, γεννιούνται ιδέες για μελλοντική έρευνα στο πεδίο, και αντιμετώπιση των βασικών του προκλήσεων. Τις ιδέες αυτές, τις συνοψίζουμε στις τρεις μεγάλες κατευθύνσεις παρακάτω:

1. *Εκτενέστερος πειραματισμός με τη ροή πληροφορίας της χειρομορφής (για προβλήματα αναγνώρισης).* Η χειρομορφή αποτέλεσε ένα πολύ σημαντικό κανάλι πληροφορίας, και εξερευνήθηκε αναλυτικά με τη χρήση τόσο διαφορετικών περιγραφητών, όσο και διαφορετικών ταξινομητών. Αν και τα αποτελέσματα στο πρόβλημα της ταξινόμησης είναι ικανοποιητικά, σε προβλήματα αναγνώρισης, τα ποσοστά επιτυχίας μειώνονται σημαντικά. Εν μέρει, δώσαμε μια εξήγηση για αυτό το αποτέλεσμα (παραπέμπουμε στην ενότητα 5.4.1), ωστόσο είναι σημαντικό να εκμεταλλευτούμε με τον πλέον εύρωστο τρόπο αυτό το κανάλι πληροφορίας, που εκτός από την ανεξάρτητη χρήση του αναμένεται να οδηγήσει σε βελτίωση του συνολικού πολυτροπικού συστήματος.

2. *Μελέτη διαφορετικών σχημάτων σύμμειξης.* Ασφαλώς είναι ιδιαίτερα σημαντικό ότι η προσέγγιση που προτείνουμε οδήγησε σε αποτελέσματα που ξεπερνούν τις επιδόσεις που επιτεύχθηκαν στο ίδιο πρόβλημα στον πρόσφατο διαγωνισμό πολυτροπικής αναγνώρισης χειρονομιών, ωστόσο είναι σημαντικό να αξιολογηθούν και διαφορετικά σχήματα σύμμειξης

τροπικοτήτων. Άλλωστε, όπως δείξαμε η προσέγγιση των "δύο περασμάτων" και ο συνδυασμός τέτοιων σχημάτων οδήγησε στη βελτίωση των επιμέρους αποτελεσμάτων, οπότε είναι εξίσου πιθανό να υπάρξει επιπλέον βελτίωση με χρήση και συνδυασμό επιπλέον σχημάτων.

3. *Περαιτέρω έρευνα στην πολυτροπική βάση χειρονομιών MOBOT*. Η βάση MOBOT που παρουσιάσαμε περιλαμβάνει ένα πολύ ενδιαφέρον υλικό για έρευνα στο πεδίο της αναγνώρισης χειρονομιών, εφόσον συγκεντρώνει μια πληθώρα πολυτροπικών, αλλά και πολυ-αισθητηριακών δεδομένων. Στα πλαίσια της παρούσας διπλωματικής έγινε μία επιλεκτική και τμηματική χρήση της, εφόσον κύριο αντικείμενο πειραματισμού αποτέλεσε η βάση ChaLearn. Ωστόσο, αποτελεί μεγάλη ερευνητική πρόκληση, να εξεταστεί το σύνολο των δεδομένων της, αλλά και να αντιμετωπιστούν οι ιδιαίτερες απαιτήσεις που παρουσιάζει.

Βιβλιογραφία

[1] A. Adjoudani and C. Benoit. On the integration of auditory and visual parameters in an HMM-based ASR. In *Speechreading by humans and machines*, pages 461–471. Springer, 1996.

[2] T. Ahmad, C. J. Taylor, and T. F. Lanitis, A. Cootes. Tracking and recognising hand gestures, using statistical shape models. *Image and Vision Computing*, 15(5):345–352, 1997.

[3] A. Argyros and M. Lourakis. Real time tracking of multiple skin-colored objects with a possibly moving camera. In *Proc. Europ. Conf. on Computer Vision*, pages 368–379, 2004.

[4] B. Bauer and K. F. Kraiss. Towards an automatic sign language recognition system using subunits. In *Proc. of Int'l Gesture Workshop*, pages 64–75, 2001.

[5] I. Bayer and T. Silbermann. A multi modal approach to gesture recognition from audio and video data. In *Proc. ACM Int'l Conf. on Multimodal Interaction*, pages 461–466, 2013.

[6] M. J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(7):828–836, 2003.

[7] H. Birk, T.B. Moeslund, and C.B. Madsen. Real-time recognition of hand alphabet gestures using principal component analysis. In *Proc. Scandinavian Conf. Image Anal.*, pages 261–268, 1997.

[8] L. Bo, K. Lai, X. Ren, and D. Fox. Object recognition with hierarchical kernel descriptors. In *Proc. Conf. on Computer Vision & Pattern Recognition*, pages 1729–1736, 2011.

[9] R. A. Bolt. Put-that-there: Voice and gesture at the graphics interface. *ACM Computer Graphics*, 14(3):262–270, 1980.

[10] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proc. of the 6th ACM international conference on Image and video retrieval.*, pages 401–408, 2007.

[11] R. Bowden and M. Sarhadi. A nonlinear model of shape and motion for tracking fingerspelt American sign language. *Image and Vision Computing*, 20(9):597–607, 2002.

[12] C.C.g Chang and C.J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[13] F. S. Chen, C. M. Fu, and C. L. Huang. Hand gesture recognition using a real-time tracking method and hidden Markov models. *Image and Vision Computing*, 21(8):745–758, 2003.

[14] Y. Chen and Y. Rui. Real-time speaker tracking using particle filter sensor fusion. *Proc. of the IEEE*, 92(3):485–494, 2004.

[15] Y. L. Chow and R. Schwartz. The n-best algorithm: An efficient procedure for finding top n sentence hypotheses. In *Proc. of the Workshop on Speech and Natural Language*, pages 199–202, 1989.

[16] G. Chrysos, E. Nikoloudakis, G. Panagiotaropoulou, G. Pavlakos, G. Retsinas, S. Theodorakis, and P. Maragos. HOG descriptor for handshape classification using Kinect sensor. In *Proc. Panhellenic Conf. Electrical and Computer Engineering Students*, pages 189–196, 2013.

[17] I. Cohen, N. Sebe, A. Garg, L.S. Chen, and T.S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1):160–187, 2003.

[18] S. Conseil, S. Bourennane, and L. Martin. Comparison of Fourier descriptors and Hu moments for hand posture recognition. In *Proc. European Conf. on Signal Processing*, pages 1960–1964, 2007.

[19] T. F. Cootes and C. J. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, 2004.

[20] Y. Cui and J. Weng. Appearance-based hand sign recognition from intensity image sequences. *Computer Vision and Image Understanding*, 78(2):157–176, 2000.

[21] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. Conf. on Computer Vision & Pattern Recognition*, pages 886–893, 2005.

[22] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proc. Europ. Conf. on Computer Vision*, pages 428–441, 2006.

[23] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongies. Eigenjoints-based action recognition using naive-Bayes-nearest-neighbor. In *2nd Joint IEEE Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[24] P. Dreuw, J. Forster, T. Deselaers, and H. Ney. Efficient approximations to model-based joint tracking and recognition of continuous sign language. In *Proc. Int'l Conf. on Automatic Face & Gesture Recognition*, pages 1–6, 2008.

[25] P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney. Benchmark databases for video-based automatic sign language recognition. In *Proc. Language Resources Evaluation Conference*, 2008.

[26] W. Du and J. Piater. Hand modeling and tracking for video-based sign language recognition by robust principal component analysis. In *Proc. ECCV Workshop on Sign, Gesture and Activity*, 2010.

[27] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, I. Guyon, V. Athitsos, H. Escalante, L. Sigal, A. Argyros, C. Sminchisescu, R. Bowden, and S. Sclaroff. Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. In *Proc. ACM Int'l Conf. on Multimodal Interaction*, pages 365–368, 2013.

[28] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athistos, and H.J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proc. ACM Int'l Conf. on Multimodal Interaction*, pages 445–452, 2013.

[29] H. Fillbrandt, S. Akyol, and K. F. Kraiss. Extraction of 3D hand shape and posture from image sequences from sign language recognition. In *Proc. Int'l Conf. on Automatic Face & Gesture Recognition*, pages 181–186, 2003.

[30] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Proc. Int'l Conf. on Computer Vision*, pages 415–422, 2011.

[31] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hammer, and H. J. Escalante. ChaLearn gesture challenge: Design and first results. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–6, 2012.

[32] J. Han, L. Shao, D. Xu, and J. Shotton. Enhanced computer vision with Microsoft Kinect sensor: A review. *IEEE Trans. on Cybernetics*, 43(5):1318–1334, 2013.

[33] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[34] M. K. Hu. Visual pattern recognition by moment invariants. *IEEE Trans. on Information Theory*, 8(2):179–187, 1962.

[35] C. L. Huang and S. H. Jeng. A model-based hand gesture recognition system. *Machine Vision and Application*, 12(5):243–258, 2001.

[36] J. M. Iverson and S. Goldin-Meadow. Why people gesture when they speak. *Nature*, 396(6708):228–228, 1998.

[37] A. Jaimes and N. Sebe. Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1):116–134, 2007.

[38] P. Kakumanu, S. Makrogiannis, and N. Bourbakis. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122, 2007.

[39] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *Proc. British Machine Vision Conference*, pages 275:1–10, 2008.

[40] D.B. Koons, C.J. Sparrell, and K.R. Thorisson. Integrating simultaneous input from speech, gaze, and hand gestures. *Intelligent Multimedia Interfaces*, pages 257–276, 1993.

[41] A. Kurakin, Z. Zhang, and Z. Liu. A real time system for dynamic hand gesture recognition with a depth sensor. In *Proc. European Conf. on Signal Processing*, pages 1975–1979, 2012.

[42] I. Laptev. On space-time interest points. *Int'l Journal of Computer Vision*, 63(2-3):107–123, 2005.

[43] I. Laptev, M. Marszałek, and C. Schmid. Learning realistic human actions from movies. In *Proc. Conf. on Computer Vision & Pattern Recognition*, pages 1–8, 2008.

[44] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. Conf. on Computer Vision & Pattern Recognition*, pages 2169–2178, 2006.

[45] H.K. Lee and J.H. Kim. An HMM-based threshold model approach for gesture recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(10):961–973, 1999.

[46] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 9–14, 2010.

[47] K. Maninis, P. Koutras, and P. Maragos. Advances on action recognition in videos using an interest point detector based on multiband spatio-temporal energies. In *Proc. Int'l Conf. on Image Processing*, 2014. (to appear).

[48] P. Maragos. Pattern spectrum and multiscale shape representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(7):701–716, 1989.

[49] M. Marszałek, I. Laptev, and S. Schmid. Actions in context. In *Proc. Conf. on Computer Vision & Pattern Recognition*, pages 2929–2936, 2009.

[50] I. Matthews and S. Baker. Active appearance models revisited. *Int'l Journal of Computer Vision*, 60(2):135–164, 2004.

[51] H. McGurk and H. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

[52] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Proc. Int'l Conf. on Computer Vision*, pages 104–111, 2009.

[53] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Trans. on Sys, Man and Cyb, Part C: Applications and Reviews*, 37(3):311–324, 2007.

[54] Y. Nam and K. Wohn. Recognition of space-time hand-gestures using hidden Markov model. In *Proc. ACM Symposium on Virtual Reality Software and Technology*, pages 51–58, 1996.

[55] K. Nandakumar, K. W. Wan, S. Chan, W. Ng, J. G. Wang, and W. Y. Yau. A multi-modal gesture recognition system using audio, video, and skeletal joint data. In *Proc. ACM Int'l Conf. on Multimodal Interaction*, pages 475–482, 2013.

[56] J.G. Neal, C.Y. Thielman, Z. Dobes, S.M. Haller, and Shapiro S.C. Natural language with integrated deictic and graphic gestures. In *Proc. of the Workshop on Speech and Natural Language*, pages 410–423, 1989.

[57] R.A. Newcombe, A. Izadi, O. Hilliges, D. Molyneaux, D. Kim, Davison A.J., P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE Int'l Symposium on Mixed and Augmented Reality*, pages 127–136, 2011.

[58] I. Oikonomidis, N. Kyriazis, and A. Argyros. Tracking the articulated motion of two strongly interacting hands. In *Proc. Conf. on Computer Vision & Pattern Recognition*, pages 1862–1869, 2012.

[59] S.C. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(6):873–891, 2005.

[60] O. Oreifej and L. Zicheng. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *Proc. Conf. on Computer Vision & Pattern Recognition*, pages 716–723, 2013.

[61] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J. R. Rohlicek. Integration of diverse recognition methodologies through reevaluation of n-best sentence hypotheses. In *Proc. of the Workshop on Speech and Natural Language*, pages 83–87, 1991.

[62] S. Oviatt and P. Cohen. Perceptual user interfaces: multimodal interfaces that process what comes naturally. *Communications of the ACM*, 43(3):45–53, 2000.

[63] X. Papageorgiou, C. Tzafestas, P. Maragos, G. Pavlakos, G. Chalvatzaki, G. Moustris, I. Kokkinos, A. Peer, B. Stanczyk, E.S. Fotinea, and E. Efthimiou. Advances in intelligent mobility assistance robot integrating multimodal sensory processing. *Universal Access in Human-Computer Interaction. Aging and Assistive Environments*, 8515:692–703, 2014.

[64] G. Pavlakos, S. Theodorakis, V. Pitsikalis, S. Katsamanis, and P. Maragos. Kinect-based multimodal gesture recognition using a two-pass fusion scheme. In *Proc. Int'l Conf. on Image Processing*, 2014. (to appear).

[65] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. of the IEEE*, 92(3):495–513, 2004.

[66] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proc. of the IEEE*, 91(9):495–513, 2003.

[67] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[68] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[69] Z. Ren, J. Yuan, J. Meng, and Z. Zhang. Robust part-based hand gesture recognition using Kinect sensor. *IEEE Trans. on Multimedia*, 15(5):1110–1120, 2013.

[70] R. C. Rose and D. B. Paul. A hidden Markov model based keyword recognition system. In *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing*, pages 129–132, 1990.

[71] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos. Dynamic affine-invariant shape-appearance handshape features and classification in sign language videos. *Journal of Machine Learning Research*, 14(1):1627–1663, 2013.

[72] B. Sapp and B. Taskar. MODEC: Multimodal decomposable models for human pose estimation. In *Proc. Conf. on Computer Vision & Pattern Recognition*, pages 3674–3681, 2013.

[73] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Proc. Conf. on Computer Vision & Pattern Recognition*, pages 32–36, 2004.

[74] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proc. of the 15th ACM Int'l Conf. on Multimedia*, pages 357–360, 2007.

[75] J. Sherrah and S. Gong. Resolving visual uncertainty and occlusion through probabilistic reasoning. In *Proc. British Machine Vision Conference*, pages 252–261, 2000.

[76] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts

from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[77] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *Proc. Europ. Conf. on Computer Vision*, pages 746–760, 2012.

[78] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden Markov models. In *Motion-Based Recognition*, pages 227–243. Springer, 1997.

[79] T. Starner, J. Weaver, and A. Pentland. Real-time American sign language recognition using desk and wearable computer based video. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[80] J. Sun, X. Wu, S. Yan, L.F. Cheong, T.S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *Proc. Conf. on Computer Vision & Pattern Recognition*, pages 2004–2011, 2009.

[81] N. Tanibata, N. Shimada, and Y. Shirai. Extraction of hand features for recognition of sign language words. In *Proc. Int'l Conf. on Vision Interface*, pages 391–398, 2002.

[82] J. Terrillon, M. Shirazi, H. Fukamachi, and S. Akamatsu. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *Proc. Int'l Conf. on Automatic Face & Gesture Recognition*, pages 54–61, 2000.

[83] P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Trans. on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.

[84] M. Turk. Multimodal interaction: A review. *Pattern Recognition Letters*, 36:189–195, 2014.

[85] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of American sign language. *Computer Vision and Image Understanding*, 81:358–384, 2001.

[86] J. Wan, Q. Ruan, W. Li, and S. Deng. One-shot learning gesture recognition from RGB-D data using bag of features. *Journal of Machine Learning Research*, 14(1):2549–2582, 2013.

[87] H. Wang, A. Kläser, C. Schmid, and C.L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *Int'l Journal of Computer Vision*, 103(1):60–79, 2013.

[88] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proc. Int'l Conf. on Computer Vision*, pages 3551–3558, 2013.

[89] H. Wang, M.M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. British Machine Vision Conference*, pages 124.1–124.11, 2009.

[90] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *Proc. Europ. Conf. on Computer Vision*, pages 872–885, 2012.

[91] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proc. Conf. on Computer Vision & Pattern Recognition*, pages 1290–1297, 2012.

[92] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proc. Europ. Conf. on Computer Vision*, pages 650–663, 2008.

[93] J. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldman. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 38(11):1870–1878, 1990.

[94] A. Wilson and A. Bobick. Parametric hidden Markov models for gesture recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(9):884–900, 1999.

[95] J. Wu, J. Cheng, C. Zhao, and H. Lu. Fusing multi-modal features for gesture recognition. In *Proc. ACM Int'l Conf. on Multimodal Interaction*, pages 453–460, 2013.

[96] Y. Wu and T. S. Huang. View-independent recognition of hand postures. In *Proc. Conf. on Computer Vision & Pattern Recognition*, volume 2, pages 88–94, 2000.

[97] Y. Wu and T.S. Huang. Vision-based gesture recognition: A review. In *Gesture-based communication in human-computer interaction*, pages 103–115. Springer, 1999.

[98] L. Xia, C.C. Chen, and J.K. Aggarwal. Human detection using depth information by Kinect. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 15–22, 2011.

[99] M. H. Yang, N. Ahuja, and M. Tabb. Extraction of 2D motion trajectories and its application to hand gesture recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(8):1061–1074, 2002.

[100] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive-Bayes-nearest-neighbor. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 14–19, 2012.

[101] X. Yang, C. Zhang, and Y. Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proc. of the 20th ACM Int'l Conf. on Multimedia*, pages 1057–1060, 2012.

[102] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013.

[103] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK book*, volume 2. Entropic Cambridge Research Laboratory Cambridge, 1997.